IMPLEMENTING GENOMIC SELECTION FOR QUANTITATIVE DISEASE RESISTANCE IN WHEAT

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Jessica Elaine Rutkoski

August 2014

IMPLEMENTING GENOMIC SELECTION FOR QUANTITATIVE DISEASE RESISTANCE IN WHEAT

Jessica Elaine Rutkoski Ph. D.

Cornell University 2014

Meeting future demands for wheat (*Triticum aestivum* L.) production will require genetic improvements in yield and yield protection against diseases. Fusarium head blight (FHB), caused by by *Fusarium graminearum,* and stem rust, caused by *Puccinia graminis* f.sp. *tritici,* are two of the most devastating diseases of wheat. Breeding for quantitative resistance (QR) to these diseases is a long-term process. Genomic selection (GS) (Haley and Visscher, 1998; Meuwissen et al., 2001), could help accelerate QR breeding. This work addresed key issues for the implementation of GS in wheat, especially for QR to FHB and stem rust. First, the importance of marker imputation and imputation method prior to GS was investigated. This is particularly important because genotyping-by-sequencing (GBS), currently the best low-cost genotyping method for wheat, results in high levels of missing data. Second, prediction models and the importance of loci-targeted genotyping for FHB and stem rust resistance were evaluated. Third, training population design strategies were explored using data from stem rust resistance breeding populations. Lastly, GS and phenotypic selection for stem rust QR were compared in terms of realized gain from selection, and impact on

inbreeding and genetic variance. The key messages of these studies are 1) missing data and choice of imputation method are not major concerns for GS as long as marker density is high, 2) in general, prediction models assuming a highly quantitative genetic architecture perform well for QR, but when loci-targeted marker data are available, a small gain in accuracy can be achieved by modeling major-effect loci more appropriately, 3) model training with historical data may require very large training population sizes and high heritabilities to achieve sufficiently high accuracies, 4) GS can be as effective as phenotypic selection, but it can lead to a faster rate of genetic variance reduction. Breeding programs implementing GS for QR and other traits should focus their efforts on how to design their breeding pipelines so that the training population can be updated often, breeding cycle time can be reduced, and effective population sizes can remain high. Choice of prediction model and marker imputation method are of lesser importance.

BIOGRAPHICAL SKETCH

Jessica Rutkoski was born in Van Nuys California on Feburary 6, 1987, to parents Marian Ritter and Dave Rutkoski. In 1992, the family moved to Wisconsin, eventually settling in the small town of Waterford, where Jessica spent her formative years and enjoyed many extracurricular activities such as working at Uncle Harry's ice cream shop, playing flute in the Milwaukee Youth Symphony Orchestra, and participating in AFS-USA as an exchange student host. In 2005, Jessica graduated high school and entered the University of Wisconsin Madison (UW-Madison). She became involved in Dr. Bill Tracy's sweet corn breeding and genetics program, where she decided to become a plant breeder. She also gained her first international experiences, visiting a high school friend in Bolivia during winter breaks and studying abroad for a semester in Argentina. In 2009 Jessica received a Bachelor of Science degree in genetics from the UW-Madison, and came to Cornell interested in addressing real world challenges facing food production. During her Ph. D. she became involved in a project focused on developing wheat germplasm with durable resistance to highly virulent races of stem rust. As part of this project, she has had the opportunity to work with and learn from scientists in in Mexico, Kenya, Ethiopia, and the United States, and has many fond memories working in the wheat fields with diverse groups of scientists. Jessica looks forward to many more such experiences and hopes to contribute her knowledge and hard work to the greater wheat community.

To my parents

ACKNOWLEDGMENTS

I would like to acknowledge all the people who contributed to this work and my development as a scientist and plant breeder. First, I would like to thank my committee members; Dr. Mark Sorrells for giving me the opportunity to become a part of his research group and for providing me with guidance, advice, and opportunities that were key to my success; Dr. Jean-Luc Jannink for his guidance and suggestions that have immensely improved this research and my education; Dr. Ronnie Coffman for providing me with the opportunity to become part of the Durable Rust Resistance in Wheat Project, which has provided funding for this research, and has enabled me to become part of the wheat community; Dr. Rebecca Nelson, for inviting me to spend time with her research group and for stimulating my interest in quantitative disease resistance and international collaboration for agricultural development. Second, I would like to thank my collaborators, especially; Dr. Ravi Singh for sharing information and germplasm and for allowing me to gain practical breeding experience; Dr. Jesse Poland for genotyping my populations, Dr. Sridhar Bhavani for managing the rust screening at Njoro, Kenya, a key location for my field experiments, and for training me on how to phenotype stem rust. I would like to thank students and post docs past and present, especially Dr. Elliot Heffner, Dr. Nicolas Heslot, Dr. Jeffrey Endelman, Dr. Deniz Akdemir, and Dr. Vahid Edriss, who have shared their guidance and knowledge through many thoughtful discussions. I would like to thank those who provided logistical support at Cornell; David Benscher, James Tanaka, and John

Shiffer, for helping me produce populations in the greenhouse, and prepare seed for internationally shipping; at Njoro Kenya, Dr. Peter Njau for managing the experiment station, Dr. Ruth Wanyera for receiving my materials, the late Samuel Kilonzo for taking care of planting and inoculations; and at Debre Zeit Ethiopia; Gebrehiwot Abraha, and Beti Hibdo who received and planted my materials for stem rust evaluation. I would also like to thank the sources of funding for this research, The Bill & Melinda Gates Foundation (Durable Rust Resistance in Wheat), United States Department of Agriculture[1]-Agricultural Research Service (USDA-ARS) Appropriation No. 5430-21000-006-00D and Hatch 149-449, USDA National Needs Fellowship Grant #2008- 38420-04755, and the American Society of Plant Biology (ASPB) -Pioneer Hi-Bred Graduate Student Fellowship. Lastly, I would like to thank my family, and loved ones for supporting me and enriching my life.

[1]Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. USDA is an equal opportunity provider and employer.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Rationale and significance

Bread wheat (*Triticum aestivum* L.) occupies more of the worlds' land area than any other cereal crop (FAO, 2012), and demand for wheat continues to rise due to population growth and increased per capita consumption (Curtis and Halford, 2014). Genetic improvement of wheat yield and yield protection, in the form of disease and pest resistance, is critical for meeting current and future demands. Fusarium head blight (FHB), predominately caused by *Fusarium graminearum*, and stem rust, caused by *Puccinia graminis* f.sp. *tritici*, are two globally important diseases of wheat (Roelfs et al., 1992; McMullen et al., 1997) that are capable of causing major losses in the regions where they occur. For example, in China, severe FHB epidemics causing up to 40% yield losses were reported between 1951 and 1985, (Zhuping, 1994), and in the United States, its estimated that FHB epidemics during the 1990s have caused three billion dollars in losses (Windels, 2000). In 1932 stem rust epidemics in eastern and central Europe lead to yield losses of 5-20% and in 1935, epidemics in North Dakota and Minnesota caused yield losses of more than 50% (Leonard and Szabo, 2005). Although effective host resistance prevented major stem epidemics for the past fifty years, with the recent emergence of a new highly virulent stem rust race group, Ug99, (Pretorius et al., 2000; Jin et al., 2008) a resurgence of severe stem

rust epidemics is imminent unless resistant varieties are deployed.

Cultivar resistance is necessary for preventing losses from FHB and stem rust. Resistance to FHB is quantitative, based on multiple genes (Buerstmayr et al., 2009), and no completely resistant varieties are available. Adequate control under high disease pressure can only be achieved by using fungicides in combination with resistant cultivars (Mesterházy et al., 2003). Resistance to stem rust can be either quantitative (Knott, 1982), or based on single genes that condition near-immunity. In regions where stem rust pathogen evolution is rapid, single resistance genes become ineffective shortly after deployment. Because quantitative resistance (QR) is generally more durable (Parlevliet, 2002; McDonald and Linde, 2002), improving stem rust QR to near- immune levels is a major goal for breeding programs targeting stem rust prone areas. Achieving adequate levels of QR to FHB or stem rust is a slow process because multiple cycles of breeding are required. A relatively new marker-assisted breeding method, genomic selection (GS) (Haley and Visscher, 1998; Meuwissen et al., 2001), has the potential to increase rates of genetic gain in crop plants (Wong and Bernardo, 2008; Lorenzana and Bernardo, 2009; Heffner et al., 2010), and could help to accelerate breeding for quantitative traits. However, there are many unknowns about the implementation of GS in wheat especially for QR improvement.

Objectives

This work addresses four major issues relevant for the implementation of GS in

wheat, and focuses primarily on stem rust QR and to a lesser extend on FHB resistance.

1. The impact of missing marker data and imputation method on GS accuracy.
2. Evaluation of GS models and the comparison of the accuracy of GS with that of marker assisted selection (MAS).
3. Choice of GS model training population.
4. Realized gain from GS compared to phenotypic selection.

Literature Review

Fusarium head blight

The most common causal organism for FHB, *Fusarium graminearum* (teliomorph= *Gibberella zea*), is a homothallic ascomycete, which can reproduce sexually and asexually during its lifecycle. The fungus overwinters as saprophytic mycelia on residues from maize (*Zea mays* L.) and small grains (Pereyra et al., Khonga and Sutton, 1988). Sexual reproduction, which does not require a sexually distinct partner, leads to the formation of ascospores that are forcibly ejected from a fruiting body called a perithecium. Disease is initiated when spores land on flowering spikletes, germinate, and enter the plant through natural openings. The fungus produces trichothecene mycotoxins including deoxynivalenol (DON), which is important for fungal spread within spikes (Bai et al., 2002). Asexual spores, conidia, are produced on infected plants and residues. The relative importance of conida and ascospores in epidemiology of the disease

is not known, but ascospores are known to be capable of long range dispersal by wind currents (Maldonado-Ramirez et al., 2005).

The two major forms of host resistance are resistance to initial infection, type I resistance, and resistance to fungal spread within a spike, type II resistance (Schroeder and Christensen, 1963). Other forms of resistance include resistance to toxin accumulation (Wang and Miller, 1988) and tolerance (Mesterhazy, 1995), which are difficult and costly to measure accurately. Breeding programs routinely evaluate FHB under field conditions after spray inoculation. Visual assessments of incidence, the percentage of infected spikes, and severity, the percentage of the spike that is infected, are used to evaluate type I and type II resistance, respectively. Genotype-by-environment interactions contribute to variation in resistance phenotypes, but more resistant cultivars are generally more stable across environments (Snijders and Van Eeuwijk, 1991; Buerstmayr et al., 2008), and although isolates of *F. graminearium* can vary in their aggressiveness, stable host-isolate interactions have not been detected (Snijders and Van Eeuwijk, 1991; Bai and Shaner, 1996).

The genetic basis of resistance to FHB is complex. Across more than forty quantitative trait loci (QTL) mapping studies, over two hundred QTLs have been detected (Liu et al., 2009; Buerstmayr et al., 2009). The most important of these is *Fhb1*, on chromosome 3B. The *Fhb1* resistance allele from the Chinese line 'Sumai-3' has been shown to reduce disease severity by 23% on average, but is not effective in all backgrounds (Pumphrey et al., 2007). *Fhb1* is also a major QTL

for DON detoxification, explaining 92.6% of the variation in a bi-parental mapping population (Lemmens et al., 2005). Although the Sumai-3 *Fhb1* allele has not been associated with significant impacts on yield and quality (Salameh et al., 2011; Tamburic-Ilincic, 2012; Bakhsh et al., 2013), it is not utilized by many breeding programs that prefer to select upon the existing variation for resistance in the 'native' germplasm. The moderate resistance found in soft red winter wheat germplasm in the eastern United States is a product of successful FHB resistance breeding using native germplasm (Sneller et al., 2010).

Stem rust

The stem rust fungus, *Puccinia graminis* f.sp. *tritici*, is a basidiomycete with a complex life cycle, reviewed in Roelfs et al. (1992) and Leonard and Szabo (2005). The asexually produced urediniospores produced on wheat are aerially dispersed, sometimes across long distances (Nagarajan and Singh, 1990), and land on a new crop of wheat plants. If conditions are favorable, spores germinate, form aspersoria over the stomata, and begin to infect the plant (Allen, 1923a; b). After successful infection, urediniospores are produced under the epidermis, causing it to rupture into a pustule, called a uredinium. These new urediniospores can again infect the same or neighboring plants, leading to multiple disease cycles in a season. Later in the season the uredinia produce teliospores where meiosis occurs. Teliospores overwinter and then germinate to produce basidiospores. The basidiospores produce mycelium that infects leaves of the alternate host barberry, *Berberis* spp., leading to the formation of

pycniospores and receptive hyphae within strcutures called pycnia. Pycniospores fuse with hyphae of the opposite mating type (Craigie, 1927; Buller, 1950), giving rise to mycelium that produces aeciospores. Aeciospores only infect wheat or other grass hosts. After germinating aeciospores penetrate wheat, uredinia and urediniospores are produced, thus completing the life cycle. In regions were barberry is not present; the fungus simply repeats the asexual cycle, surviving on living grass hosts.

Like other rust fungi, *P. graminis* interacts with its host in a gene-for-gene manner (Flor, 1971). In general, a single major-effect resistance gene (R-gene) encodes a protein that can recognize a single fungal effector (Jones and Dangl, 2006). The majoritiy of R-genes that have been characterized encode nucleotide binding leucine-rich repeat proteins. The leucine-rich repeat domain recognizes the a specific fungal effector, and the nucleotide binding domain interacts with downstream signaling molecules(Collier and Moffett, 2009). Recognition of the fungus leads to a hypersensitive response (Stakman, 1915), preventing spread of the pathogen. *P. graminis* races that can evade recognition by the hosts' R-genes are selected in the population, eventually rendering these R-genes ineffective. Races of stem rust (Stakman et al., 1962) are named according to the R-genes they can overcome. The appearance of new *P. graminis* races motivates the continual search for more R-genes. Currently, there are are over fifty known stem rust R-genes in wheat (McIntosh et al., 1995). Many have been used extensively in breeding, and in some regions stem rust is very effectively managed through R-

gene deployment and race surveillance (Knott, 1972; Park, 2008). Where evolution of *P. graminis* occurs rapidly due to favorable environmental conditions, continual cropping of wheat, and/or abundance of barberry, deployment of QR, also called adult plant resistance, is advocated because of its durability.

Stem rust QR is known to be associated with several additive loci (Yu et al., 2011; Njau et al., 2012; Singh et al., 2013) including *Sr2* on chromosome 3B, (Sunderwirth and Roelfs 1980) and *Sr57*/*Lr34* on chromosome 7D (Kerber and Green, 1980; Kerber and Aung, 1999; Vanegas et al., 2008; Rouse et al., 2014). *Sr2* was transferred to wheat from tetraploid emmer (McFadden, 1930), and is present Australian, North American, and CIMMYT germplasm (Mago et al., 2011). *Sr57*/*Lr34* originates from Asia and is distributed globally (Dakouri et al., 2013). Both *Sr2* and *Sr57*/*Lr34* have been used widely in agriculture and have remained effective for over fifty years (McIntosh et al., 1995). It is thought that these genes and other unknown stem rust QR genes are not involved in host-pathogen recognition, and do not interact with the pathogen in a gene-for-gene manner. *Sr57*/*Lr34,* the only rust QR gene cloned, encodes a putative ABC transporter (Krattinger et al., 2009), suggesting that its mechanism of action of different than that of R-genes. Currently, only *Sr2* and *Sr57*/*Lr34* are being used in MAS for QR. MAS or GS could be especially useful to select for QR in the presence of R-genes which inhibit breeding for QR because they can completely mask the underlying QR phenotype (van der Plank, 1963).

Genomic selection

Excitement about using MAS to improve breeding efficiency began more than twenty years ago (Tanksley, S. D Young N.D., Paterson A. H., 1989). Since then, numerous MAS strategies have been developed, including marker assisted backcrossing (Tanksley, 1983) with foreground and background selection (Hillel et al., 1990), enrichment of favorable alleles in early generations (Howes and Woods, 1998; Bonnett et al., 2005), selection for quantitative traits using markers at multiple loci (Fernando and Grossman, 1989; Lande and Thompson, 1990) across multiple cycles of selection (Zhang and Smith, 1992), and finally, GS (Haley and Visscher, 1998; Meuwissen et al., 2001), currently the most effective MAS strategy for quantitative traits (Bernardo and Yu, 2007; Massman et al., 2013). Due to high costs of genotyping, backcross introgression of major-effect disease resistance alleles has been the most successful MAS technique in plant breeding (Bernardo, 2008; Xu and Crouch, 2008), especially in the public sector. However, as the cost of sequencing declines (Wetterstrand, 2014), the possibility of improving quantitative traits more efficiently with GS becomes more realistic.

GS is the selection of individuals using genome-wide marker based predictions of breeding value. To perform GS, a population that has been both genotyped and phenotyped, referred to as the training population (TP), is required. The TP is used to train or calibrate a statistical model that can then be used to predict breeding values of selection candidates that have not yet been phenotyped. Selections of new breeding parents are made based on these

predictions, leading to shorter breeding cycles. To maintain prediction model accuracy, some of the selection candidates that were targets for GS are phenotyped, and this information is used to update the model. The expected gain from GS per unit time is defined as $\Delta G = ir\sigma_A/T$, where i is the selection intensity, r is the selection accuracy, σ_A is the square root of the additive genetic variance, and T is the length of time to complete one breeding cycle (Falconer and Mackay, 1996, p. 189). Assuming equal selection intensities and equal genetic variance for both GS and PS, GS can lead to greater gain per unit time as long as long as the reduction in in breeding cycle duration from GS more than compensates for the reduction in selection accuracy. Given realistic assumptions of selection accuracies, breeding cycle times, and selection intensities, GS has been shown to enable increased gain per unit time compared to PS in both animal and crop breeding (Heffner et al., Schaeffer, 2006; Wong and Bernardo, 2008; Lorenzana and Bernardo, 2009; Zhong et al., 2009)

The genetic architecture of the trait of interest is key for determining the utility of GS compared to conventional MAS. While GS is the best marker assisted-breeding strategy for quantitative traits, for Mendelian traits, conventional MAS would be the best strategy to maximize genetic gain per unit time and cost. For traits such as QR to stem rust and FHB that are conferred by some large-effect and small-effect QTL, it is unclear whether GS would be more effective than MAS. Furthermore, simulation studies have shown that genetic architecture should affect the relative performance of different GS models (Daetwyler et al., 2010;

Wimmer et al., 2013). For example, Bayesian methods that treat markers heterogenously should perform better when there are fewer QTL compared to Ridge-Regression or Genomic Best Linear Unbiased Prediction (Daetwyler et al., 2010). In spite of the theoretical advantages of Bayesian methods for oligogenic traits, most emperical studies have found Bayesian methods perform approximately as well Ridge-Regression or Genomic Best Linear Unbiased Prediction across a wide range of traits (Hayes et al., 2009a; Heffner et al., 2011; Heslot et al., 2012; Wimmer et al., 2013).

Improving the accuracy of GS has been the focus of many studies. Assuming markers and QTL are in perfect linkage disequilibrium, accuracy is determined by the TP size (N), heritability of the trait (h^2) in the TP, and the effective number of loci Me:

$$r = \sqrt{\frac{Nh^2}{Nh^2 + Me}}$$

(Goddard, 2009; Daetwyler et al., 2010). When QTL and markers are in imperfect linkage disequilibrium, accuracy will be lower unless the TP and selection candidates are closely related (de Los Campos et al., 2013). Thus, in realistic scenarios, the relationship between the TP and the selection candidates is a key factor affecting accuracy (Habier et al., 2007; de Roos et al., 2009; Hayes et al., 2009b; Long et al., 2011; Pszczola et al., 2012). Other factors that can sometimes affect accuracies include genotype-by-environment interaction between the TP and breeding target environments (Ly et al., 2013; Dawson et al., 2013), choice of

statistical model (Heslot et al., 2012), marker platform (Solberg et al., 2008; Poland et al., 2012), and genotype imputation method (Rutkoski et al., 2013).

With so many variables to consider, it becomes nearly impossible to predict the accuracy of GS in advance. Empirical validation studies are necessary to determine how GS could perform in a particular set of germplasm for the trait(s) of interest. Validation studies can also be used to develop recommendations within breeding programs about choice of prediction model, imputation method, TP size, and TP composition. However, ultimately GS strategies will need to be tested 'the hard way' by putting them into practice in breeding programs.

Dissertation organization

The second chapter addresses the impact of missing marker data and imputation method on GS accuracy in general using datasets from maize, wheat and barley. The third and fourth chapters focus on evaluating GS prediction methods and genotyping strategies for FHB and stem rust QR, respectively. Both include a comparison of GS and marker assisted selection accuracy. The fourth chapter also determines if accuracy can be gained by modeling large effect QTL as fixed effects. The fifth chapter addresses choice of training population composition under different training population size, and heritability scenarios. The sixth chapter is an evaluation of realized gain from GS for stem rust QR in comparison with phenotypic selection. The seventh and final chapter provides a summary and highlights the overall conclusions of the dissertation.

References

Allen, R.F. 1923a. A cytological study of infection of Baart and Kanred wheats by Puccinia graminis tritici. J. Agric. Res. 23: 131–152.

Allen, R.F. 1923b. Cytological studies of infection of Baart, Kanred, and Mindum wheats by Puccina graminis tritici forms III and XIX. J. Agric. Res. 33: 201–222.

Bai, G.H., A.E. Desjardins, and R.D. Plattner. 2002. Deoxynivalenol-nonproducing *Fusarium graminearum* causes initial infection, but does not cause disease spread in wheat spikes. Mycopathologia 153: 91–98

Bai, G.-H., and G. Shaner. 1996. Variation in *Fusarium graminearum* and cultivar resistance to wheat scab. Plant Dis. 80: 975–979.

Bakhsh, A., N. Mengistu, P.S. Baenziger, I. Dweikat, S.N. Wegulo, D.J. Rose, G. Bai, and K.M. Eskridge. 2013. Effect of Fusarium head blight resistance gene on agronomic and end-use quality traits of hard red winter wheat. Crop Sci. 53: 793-801.

Bernardo, R. 2008. Molecular markers and selection for complex traits in plants: Learning from the last 20 years. Crop Sci. 48: 1649–1664.

Bernardo, R., and J. Yu. 2007. Prospects for genome-wide selection for quantitative traits in maize. Crop Sci. 47: 1082–1090.

Bonnett, D.G., G.J. Rebetzke, and W. Spielmeyer. 2005. Strategies for efficient implementation of molecular markers in wheat breeding. Mol. Breed. 15: 75–85.

Buerstmayr, H., T. Ban, and J.A. Anderson. 2009. QTL mapping and marker-assisted selection for Fusarium head blight resistance in wheat: a review. Plant Breed. 128: 1–26.

Buerstmayr, H., M. Lemmens, M. Schmolke, G. Zimmermann, L. Hartl, F. Mascher, M. Trottet, N.E. Gosman, and P. Nicholson. 2008. Multi-environment evaluation of level and stability of FHB resistance among parental lines and selected offspring derived from several European winter wheat mapping populations. Plant Breed. 127: 325–332.

Buller, A.H.R. 1950. Researches on fungi Vol. 7: The sexual process in the Uredinales. University of Tronto, Toronto, Canada.

Collier, S.M., and P. Moffett. 2009. NB-LRRs work a “bait and switch” on pathogens. Trends Plant Sci. 14: 521–529.

Craigie, J.H. 1927. Discovery of the Function of the Pycnia of the Rust Fungi. Nature 120: 765–767.

Curtis, T., and N.G. Halford. 2014. Food security: the challenge of increasing wheat yield and the importance of not compromising food safety. Ann. Appl. Biol. 164: 354–372.

Daetwyler, H.D., R. Pong-Wong, B. Villanueva, and J.A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185: 1021–1031.

Dakouri, A., B.D. McCallum, and S. Cloutier. 2013. Haplotype diversity and evolutionary history of the Lr34 locus of wheat. Mol. Breed. 33: 639–655.

Dawson, J.C., J.B. Endelman, N. Heslot, J. Crossa, J. Poland, S. Dreisigacker, Y. Manès, M.E. Sorrells, and J.-L. Jannink. 2013. The use of unbalanced historical data for genomic selection in an international wheat breeding program. F. Crop. Res. 154: 12–22.

Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to quantitative genetics. 4th ed. Longman, New York, NY.

Fernando, R.L., and M. Grossman. 1989. Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. 21: 246–477.

Flor, H.H. 1971. Current Status of Gene-For-Gene Concept. Annu. Rev. Phytopathol. 9: 275–296.

Food and Agriculture of the United Nations, FAOSTAT database. 2012. Available at http://faostat.fao.org/site/291/default.aspx (verified 5 June 2014).

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136: 245–257.

Habier, D., R.L. Fernando, and J.C.M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177: 2389–2397.

Haley, C.S., and P.M. Visscher. 1998. Strategies to utilize marker-quantitative trait loci associations. J. Dairy Sci. 81: 85–97.

Hayes, B.J., P.J. Bowman, a J. Chamberlain, and M.E. Goddard. 2009a. Invited review: Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92: 433–443.

Hayes, B.J., P.M. Visscher, and M.E. Goddard. 2009b. Increased accuracy of selection by using the realized relationship matrix. Genet. Res. (Camb). 91: 47–60.

Heffner, E.L., J.-L. Jannink, and M.E. Sorrells. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Gen. 4: 1–11.

Heffner, E.L., A.J. Lorenz, J.-L. Jannink, and M.E. Sorrells. 2010. Plant breeding with genomic selection: gain per unit time and cost. Crop Sci. 50: 1681–1690.

Heslot, N., H.-P. Yang, M.E. Sorrells, and J.-L. Jannink. 2012. Genomic selection in plant breeding: a comparison of models. Crop Sci. 52: 146-160.

Hillel, J., T. Schaap, A. Haberfeld, A.J. Jeffreys, Y. Plotzky, A. Cahaner, and U. Lavu. 1990. DNA fingerprints applied to gene introgression in breeding programs. Genetics 789: 783–789.

Howes, N.K., and S.M. Woods. 1998. Simulations and practical problems of applying multiple marker assisted selection and doubled haploids to wheat breeding programs. Eur. J. Plant Pathol. 100: 225–230.

Jin, Y., L.J. Szabo, Z.A. Pretorius, R.P. Singh, R. Ward, and T. Fetch. 2008. Detection of virulence to resistance gene Sr24 within race TTKS of *Puccinia graminis* f. sp *tritici*. Plant Dis. 92(6): 923–926.

Jones, J.D.G., and J.L. Dangl. 2006. The plant immune system. Nature 444(7117): 323–329.

Kerber, E.R., and T. Aung. 1999. Leaf rust resistance gene lr34 associated with nonsuppression of stem rust resistance in the wheat cultivar Canthatch. Phytopathology 89: 518–521.

Kerber, E.R., and J. Green. 1980. Suppression of stem rust resistance in the hexaploid wheat cv. Canthatch by chromosome 7DL1. Can. J. Bot. 58: 1347–1350.

Khonga, E.B., and J.C. Sutton. 1988. Inoculum production and survival of *Gibberella zeae* in maize and wheat residues. Can. J. Plant Pathol. 10: 232–239.

Knott, D.R. 1972. Using race-specific resistance to manage the evolution of plant pathogens. J. Environ. Qual. 1: 227–231.

Knott, D.R. 1982. Multigenic inheritance of stem rust resistance in wheat. Crop Sci. 22: 393–399.

Krattinger, S.G., E.S. Lagudah, W. Spielmeyer, R.P. Singh, J. Huerta-Espino, H. McFadden, E. Bossolini, L.L. Seiter, B. Keller, and L.L. Selter. 2009. A putative ABC transporter confers durable resistance to multiple fungal pathogens in wheat. Sci. 323: 1360–1363.

Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124: 743–756.

Lemmens, M., U. Scholz, F. Berthiller, C.D. Asta, A. Koutnik, R. Schuhmacher, G. Adam, H. Buerstmayr, Á. Mesterházy, R. Krska, and P. Ruckenbauer. 2005. The ability to detoxify the mycotoxin deoxynivalenol colocalizes with a major quantitative trait locus for fusarium head blight resistance in wheat. 18: 1318–1324.

Leonard, K.L., and L.J. Szabo. 2005. Pathogen profile: Stem rust of small grains and grasses caused by *Puccinia graminis*. Mol. Plant Pathol. 6: 99–111.

Liu, S., M.D. Hall, C.A. Griffey, and A.L. McKendry. 2009. Meta-analysis of QTL associated with Fusarium head blight resistance in wheat. Crop Sci. 49: 1955–1968.

Long, N., D. Gianola, G.J.M. Rosa, and K.A. Weigel. 2011. Long-term impacts of genome-enabled selection. J. Appl. Genet. 52: 467–480.

Lorenzana, R.E., and R. Bernardo. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. 120: 151–161.

de Los Campos, G., A.I. Vazquez, R. Fernando, Y.C. Klimentidis, and D. Sorensen. 2013. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. 9: e1003608.

Ly, D., M.T. Hamblin, I. Rabbi, G. Melaku, M. Bakare, H.G. Gauch, R. Okechukwu, A.G.O. Dixon, P. Kulakow, and J.-L. Jannink. 2013. Relatedness and genotype-by-environment interaction affect prediction accuracies in genomic selection: A study in cassava. Crop Sci. 53: 1312–1325.

Mago, R., G. Brown-Guedira, S. Dreisigacker, J. Breen, Y. Jin, R. Singh, R. Appels, E.S. Lagudah, J. Ellis, and W. Spielmeyer. 2011. An accurate DNA marker assay for stem rust resistance gene Sr2 in wheat. Theor. Appl. Genet. 122: 735–744.

Maldonado-Ramirez, S.L., D.G. Schmale, E.J. Shields, and G.C. Bergstrom. 2005. The relative abundance of viable spores of *Gibberella zeae* in the planetary boundary layer suggests the role of long-distance transport in regional epidemics of Fusarium head blight. Agric. For. Meteorol. 132: 20–27.

Massman, J.M., H.-J.G. Jung, and R. Bernardo. 2013. Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. Crop Sci. 53: 58–66.

McDonald, B. A, and C. Linde. 2002. Pathogen population genetics, evolutionary potential, and durable resistance. Annu. Rev. Phytopathol. 40: 349–379.

McFadden, E.S. 1930. A successful transfer of emmer characters to vulgare wheat. J. Am. Soc. Agron. 22: 1020–1034.

McIntosh, R.A., C.R. Wellings, and R.F. Park. 1995. Wheat rusts: an atlas of resistance genes. Kluwer Academic Publishers.

McMullen, M., R. Jones, and D. Gallenberg. 1997. Scab of wheat and barley: a re-emerging disease of devastating impact. Plant Dis. 81: 1340–1348.

Mesterhazy, A. 1995. Types and components of resistance to Fusarium head blight of wheat. Plant Breed. 114: 377–386.

Mesterházy, Á., T. Bartók, C. Lamper, N. Company, and P.O. Box. 2003. Influence of Wheat Cultivar , Species of Fusarium , and Isolate Aggressiveness on the Efficacy of Fungicides for Control of Fusarium Head Blight. Plant Dis. 87: 1107–1115.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Nagarajan, S., and D.V. Singh. 1990. Long-distance dispersion of rust pathogens. Annu. Rev. Phytopathol. 28: 139–153.

Njau, P.N., S. Bhavani, J. Huerta-Espino, B. Keller, and R.P. Singh. 2012. Identification of QTL associated with durable adult plant resistance to stem rust race Ug99 in wheat cultivar “Pavon 76.” Euphytica 190: 33–44.

Park, R.F. 2008. Breeding cereals for rust resistance in Australia. Plant Pathol. 57: 591–602.

Parlevliet, J.E. 2002. Durability of resistance against fungal, bacterial and viral pathogens; present situation. Euphytica 124: 147–156.

Pereyra, S.A., R. Dill-Macky, and A.L. Sims. Survival and inoculum production of *Gibberella zeae* in wheat residue. Plant Dis. 88: 724–730.

Van der Plank, J.E. 1963. Plant diseases: epidemics and control. Academic Press, New York, NY.

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sánchez-Villeda, M. Sorrells, and J.-L. Jannink. 2012. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. Plant Gen. 5: 103–113.

Pretorius, Z.A., R.P. Singh, W.W. Wagoire, and T.S. Payne. 2000. Detection of virulence to wheat stem rust resistance gene Sr31 in *Puccinia graminis* f. sp. *tritici* in Uganda. Plant Dis. 84: 203.

Pszczola, M., T. Strabel, H.A. Mulder, and M.P.L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95: 389–400.

Pumphrey, M.O., R. Bernardo, and J.A. Anderson. 2007. Validating the Fhb1 QTL for Fusarium head blight resistance in near-isogenic wheat lines developed from breeding populations. Crop Sci. 47: 200–206.

Roelfs, A.P., R.P. Singh, and E.E. Saari. 1992. Rust diseases of wheat: Concepts and methods of disease management. CIMMYT, Mexico, D.F.

De Roos, A.P.W., B.J. Hayes, and M.E. Goddard. 2009. Reliability of genomic predictions across multiple populations. Genetics 183(4): 1545–1553.

Rouse, M.N., L.E. Talbert, D. Singh, and J.D. Sherman. 2014. Complementary epistasis involving Sr12 explains adult plant resistance to stem rust in Thatcher wheat (*Triticum aestivum* L.). Theor. Appl. Genet. 128:1-11

Rutkoski, J.E., J. Poland, J.-L. Jannink, and M.E. Sorrells. 2013. Imputation of unordered markers and the impact on genomic selection accuracy. G3 (Bethesda). 3: 427–439.

Salameh, A., M. Buerstmayr, B. Steiner, A. Neumayer, M. Lemmens, and H.

Buerstmayr. 2011. Effects of introgression of two QTL for Fusarium head blight resistance from Asian spring wheat by marker-assisted backcrossing into European winter wheat on Fusarium head blight resistance, yield and quality traits. Mol. Breed. 28: 485–494.

Schaeffer, L.R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123: 218–223.

Schroeder, H.W., and J.J. Christensen. 1963. Factors affecting resistance of Wheat to scab caused by *Gibberella zeae*. Phytopathology 53: 831–838.

Singh, S., R.P. Singh, S. Bhavani, J. Huerta-Espino, and E.E. Lopez-Vera. 2013. QTL mapping of slow-rusting, adult plant resistance to race Ug99 of stem rust fungus in PBW343/Muu RIL population. Theor. Appl. Genet. 126: 1367–1375.

Sneller, C.H., P. Paul, and M. Guttieri. 2010. Characterization of resistance to Fusarium head blight in an eastern U.S. soft red winter wheat population. Crop Sci. 50: 123–133.

Snijders, C.H., and F. a Van Eeuwijk. 1991. Genotype x strain interactions for resistance to Fusarium head blight caused by *Fusarium culmorum* in winter wheat. Theor. Appl. Genet. 81: 239–244.

Solberg, T.R., A.K. Sonesson, J.A. Woolliams, and T.H.E. Meuwissen. 2008. Genomic selection using different marker types and densities. 86: 2447–2454.

Stakman, E.C. 1915. Relation between *Puccinia graminis* and plants highly resistant to its attack. J. Agric. Res. 4: 193–299.

Stakman, E.C., D.M. Steward, and W.Q. Loegering. 1962. Identification of physiologic races of *Puccinia graminis* var. *tritici*. U.S. Dep. Agric. Res. Serv. E-617.

Sunderwirth, S.D., and A.P. Roelfs. 1980. Greenhouse evaluation of the adult-plant resistance of Sr2 to wheat-stem rust. Phytopathology 70: 634–637.

Tamburic-Ilincic, L. 2012. Effect of 3B, 5A and 3A QTL for Fusarium head blight resistance on agronomic and quality performance of Canadian winter wheat (T Miedaner, Ed.). Plant Breed. 131: 722–727.

Tanksley, S.D. 1983. Molecular markers in plant breeding. Plant Mol. Biol. Report. 1: 3–8.

Tanksley, S. D Young N.D., Paterson A. H., B.M.W. 1989. RFLP Mapping in plant breeding: new tools for and old science. Nature 7: 257–264.

Vanegas, C.D.G., D.F. Garvin, and J. a. Kolmer. 2008. Genetics of stem rust resistance in the spring wheat cultivar Thatcher and the enhancement of stem rust resistance by Lr34. Euphytica 159: 391–401.

Wang, Y.Z., and J.D. Miller. 1988. Effects of *Fusarium graminearum* metabolites on wheat tissue in relation to Fusarium Head Blight Resistance. J. Phytopathol. 122: 118–125.

Wetterstrand, K.A. 2014. DNA sequencing costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: www.genome.gov/sequencingcosts (verified 12 June 2014).

Wimmer, V., C. Lehermeier, T. Albrecht, H.-J. Auinger, Y. Wang, and C.-C. Schön. 2013. Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics 195: 573–587.

Windels, C.E. 2000. Economic and social impacts of Fusarium head blight: changing farms and rural communities in the northern great plains. Phytopathology 90: 17–21.

Wong, C.K., and R. Bernardo. 2008. Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theor. Appl. Genet. 116: 815–824.

Xu, Y., and J.H. Crouch. 2008. Marker-assisted selection in plant breeding: from publications to practice. Crop Sci. 48: 391–407.

Yu, L.-X., A. Lorenz, J. Rutkoski, R.P. Singh, S. Bhavani, J. Huerta-Espino, and M.E. Sorrells. 2011. Association mapping and gene-gene interaction for stem rust resistance in CIMMYT spring wheat germplasm. Theor. Appl. Genet. 123: 1257–1268.

Zhang, W., and C. Smith. 1992. Computer simulation of marker-assisted selection utilizing linkage disequilibrium. Theor. Appl. Genet. 83: 813–820.

Zhong, S., J.C.M. Dekkers, R.L. Fernando, and J.-L. Jannink. 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. Genetics 182: 355–64.

Zhuping, Y. 1994. Breeding for resistance to Fusarium head blight of wheat in the Mid-Lower Yangtze River Valley of China. Wheat special report no. 27.

Mexico, D.F.

CHAPTER 2

IMPUTATION OF UNORDERED MARKERS AND THE IMPACT ON GENOMIC SELECTION ACCURACY[2]

Abstract

Genomic selection (GS), a breeding method that promises to accelerate rates of genetic gain, requires dense, genome-wide marker data. Genotyping-by-sequencing can generate a large number of *de novo* markers. However, without a reference genome, these markers are unordered and typically have a large proportion of missing data. Because marker imputation algorithms were developed for species with a reference genome, algorithms suited for unordered markers have not been rigorously evaluated. Using four empirical datasets, we evaluate and characterize four such imputation methods referred to as k-nearest neighbors, singular value decomposition, random forest regression, and expectation maximization imputation in terms of their imputation accuracies and the factors affecting accuracy. The effect of imputation method on the GS accuracy is assessed in comparison with mean imputation. The effect of excluding markers with a large proportion of missing data on the GS accuracy is also examined. Our results show that imputation of unordered markers can be

---

[2] Originally published as: Rutkoski, J.E., J. Poland, J.-L. Jannink, & M.E. Sorrells, 2013. Imputation of unordered markers and the impact on genomic selection accuracy. G3 (Bethesda, Md.), 3: 427–439.

accurate especially when linkage disequilibrium between markers is high, and genotyped individuals are related. Of the methods evaluated, random forest regression imputation produced superior accuracy. In comparison with mean imputation, all four imputation methods we evaluated led to higher GS accuracies when the level of missing data was high. Including rather than excluding markers with a large proportion of missing data nearly always led to greater GS accuracies. We conclude that high levels of missing data in dense marker sets is not a major obstacle for GS, even when marker order is not known.

Abbreviations

GS, genomic selection; SC, selection candidate; GEBV, genomic estimated breeding value; LD, linkage disequilibrium; GBS, genotyping-by-sequencing; SNP, single-nucleotide polymorphism; MNI, mean imputation; kNNI, k-nearest neighbors imputation; SVDI, singular value decomposition imputation; EMI, expectation maximization imputation; RFI, random forest regression imputation; WW, winter wheat; SW, spring wheat; DTM, drought tolerant maize; NAB, North American barley; SRRW, stem rust resistant wheat; DArT, Diversity Array Technology; HT, height; DTH, days to heading; CIMMYT, International Maize and Wheat Improvement Center; B-glucan, beta-glucan; CPU, central processing unit; MAF, minor allele frequency; PEV, prediction error variance

Introduction

Genomic selection (GS) (Meuwissen et al., 2001) is a relatively new breeding methodology reviewed by Hayes et al. (2009), Heffner et al. (2009) and

Lorenz et al. (2011) that is increasingly attractive for the genetic improvement of various species because of its potential to increase the rate of genetic gain (Wong and Bernardo, 2008; Lorenzana and Bernardo, 2009; Heffner et al., 2010). With GS, a training population having both phenotypic data and genome-wide marker data is used to develop a prediction model for the trait of interest. Prior to phenotyping, this prediction model is then applied to selection candidates (SC)s that have been genotyped. Genomic estimated breeding values (GEBVs) are calculated for the SCs and selections are made using these values. These breeding values are estimated using genotypes instead of phenotypes, therefore, selection can occur in early stages on a single plant basis or in situations where phenotyping is either not possible, unreliable, or too expensive, thus leading to shorter selection cycles.

One of the requirements for GS is genome-wide marker coverage. In general, one marker should be in linkage disequilibrium (LD) with each segregating segment of the genome. The choice of marker platform is driven by the available genotyping technology and the cost per data-point. Genotyping-by-sequencing (GBS) is gaining popularity because it can be less expensive than other platforms and can provide genome-wide marker coverage for species that lack genotyping resources such as pre-designed single-nucleotide polymorphism (SNP) platforms (Poland and Rife, 2012). Polymorphic loci scored by GBS can contain a large proportion of missing data across samples because random fragments of the genome are sequenced at low depth, leading some loci to have

zero coverage in some individuals (Elshire et al., 2011). The proportion of missing data depends on the sequencing depth and library complexity. Greater sequencing depth leads to a smaller proportion of missing data but increases genotyping cost. Less complex libraries on the other hand will have less missing data but a fewer markers. In order to generate a large number of markers at low cost, low sequencing depth is commonly used, leading to a large proportion of missing data points. Most analyses require a complete dataset; therefore, marker imputation is a necessary step before GBS data can be used for most purposes.

Imputation has been shown to increase power in association mapping studies (Marchini et al., 2007; Marchini and Howie, 2010) and, for GS, imputation can enable the use of low-density genotyping without a major loss in accuracy because a closely related reference panel genotyped at high density can be used to impute markers not present in the low-density marker panel. (Habier et al., 2009; Weigel et al., 2010; Dassonneville et al., 2011; Mulder et al., 2012). Although several highly accurate and widely used imputation algorithms have been developed to assign allelic states of missing values in genotype data, reviewed by Pei et al. (2008) and Marchini et al. (2010), these algorithms were designed for human genetic data and they require that the order of the markers be known because they are based on constructing haplotypes. For species lacking a reference genome and complete reference linkage map such as wheat, *Triticum aestivum* L.*,* the majority of markers typed on a given population are unordered and current genotype imputation methods cannot be used. Although

for bi-parental populations linkage maps can be constructed, breeding populations for genomic selection are derived from multiple parents and not well structured for developing genetic maps. Thus, alternative imputation strategies that are map-independent are necessary when GBS is used for species lacking a reference genome sequence and for populations unsuitable for linkage map construction. There are many general imputation methods that do not require any prior information about the variables to be imputed. Although these methods are used across many disciplines, they have not been tested for imputation accuracy of genome-wide marker data. It is also not known how imputation with a general and potentially less accurate method prior to GS model training will affect the GS model accuracy. However, we expect these imputation methods to improve the GS accuracy because during the imputation step, genotypic information from both the training and selection sets is used to estimate missing values. Thus, the validation set helps improve imputation of the training set and vice versa.

The objective of this study was to evaluate imputation strategies that do not require prior information about the order of the markers. The imputation methods compared were: mean imputation (MNI), k-nearest neighbors imputation (kNNI) (Troyanskaya et al., 2001), singular value decomposition imputation (SVDI) (Troyanskaya et al., 2001), expectation maximization imputation (EMI) (Dempster et al., 1977), and random forest regression imputation (RFI) (Stekhoven and Bühlmann, 2011). Using array-based genotypic

datasets with varying levels of simulated missing data, these methods were compared in terms of their imputation accuracy, computation time, and impact on GS prediction accuracy. The factors affecting imputation accuracy for each method at the marker genotype and individual genotype level were also examined. Lastly, we determine if excluding rather than including markers with high levels of missing data could lead to higher accuracy.

Materials and methods

Original datasets

We used five different datasets consisting of genome-wide markers and breeding value estimates. These datasets are referred to as winter wheat (WW), spring wheat (SW), drought tolerant maize (DTM), North American barley (NAB), and stem rust resistant wheat (SRRW). The WW data consists of 374 elite inbred individuals originating from the Cornell winter wheat breeding program. The markers consisted of 1158 polymorphic diversity array technology (DArT) (Akbari et al., 2006) markers coded as "-1", and "1". For a more detailed description of this dataset refer to Heffner et al. (2011). The traits used for the evaluation of cross-validated GS accuracies for WW were grain yield, height (HT), protein, and days to heading (DTH). The SW data is a historical dataset consisting of 599 elite inbred spring wheat lines originating from the International Maize and Wheat Improvement Center (CIMMYT) wheat breeding program. The markers consist of 1279 polymorphic DArT markers coded as "0" and "1" and the trait used for the evaluation of cross-validated GS accuracies was grain yield in

CIMMYT mega-environment 1. The DTM data consists of 264 tropical CIMMYT maize lines. The trait used to calculate cross-validated GS model accuracies for DTM was grain yield. The marker data consists of SNPs coded as "-1", "0", and "1". For more details about the SW and DTM datasets, or to access these datasets, refer to Crossa et al. (2010). The NAB data set consists of a North American spring barley association mapping panel evaluated from 2006-2008 as part of the Barley Coordinated Agricultural Project (2011). The panel consists of 911 individuals with 2146 polymorphic SNPs. The trait used to calculate GS model accuracies was beta-glucan content (B-glucan). The data can be accessed at http://triticeaetoolbox.org/barley.

The SRRW data set consists of 360 recent, elite CIMMYT spring wheat lines that have been selected for quantitative resistance to stem rust caused by *Puccinia graminis* f.sp. *tritici*. The markers consist of over 130,000 GBS polymorphisms. Three different versions of the SRRW GBS data, described in Table 2.1, were created based on different per-marker percent missing data thresholds. For the first version referred to as SRRW version NA20, markers were excluded if they had more than 20% missing values, which resulted in 2,014 total markers. For the second set and third sets, referred to as SRRW versions NA50 and NA70, markers were excluded if they had more than 50% and 70% missing data, respectively, and then 2,014 markers were randomly selected. The percent of the data points that were missing in the original WW, SW, DTM, and

Table 2.1: Description of datasets used for imputation and genomic selection

Dataset†	Version‡	Mean percent missing data points§	Number of markers	Number of individuals
WW	NA20	12.13	1158	374
	NA50	34.08	1158	374
	NA70	58.84	1158	374
SW	NA20	12.1	1279	599
	NA50	34.98	1279	599
	NA70	60.54	1279	599
DTM	NA20	11.99	1135	264
	NA50	34.9	1135	264
	NA70	60.53	1135	264
NAB	NA20	12.1	2146	911
	NA50	35.03	2146	911
	NA70	60.49	2146	911
SRRW	NA20	12.16	2014	360
	NA50	35.13	2014	360
	NA70	60.72	2014	360

†WW: Cornell winter wheat, SW: CIMMYT elite spring wheat, DTM: CIMMYT drought tolerant maize, NAB: North American barley, SRRW: CIMMYT stem rust resistant wheat
‡NA20: up to 20% missing data per marker, NA50: up to 50% missing data per marker, NA70: up to 70% missing data per marker
§The percent of total data points that are missing

NAB datasets was between 0.2-3%. This low level of pre-existing missing data was assumed to have a negligible effect on the imputation and GS accuracies and for these datasets the original marker data is referred to as version NA0.

Calculation of linkage disequilibrium between marker pairs

For the original WW, SW. DTM, and NAB datasets, LD between all marker pairs was measured using the r^2 statistic, where r^2 between two markers was calculated using the formula:

$$r^2 = \frac{D^2}{p_1 q_1 p_2 q_2}$$

where $D = x_{11} - p_1 p_2$ is the probability of observing the combination of allele 1 at marker *j* and allele 1 at marker *l*, p_1 is the probability of allele 1 at marker *j*, q_1 is the probability of allele 2 at marker *j*, p_2 is the probability of allele 1 at marker *l*, and q_2 is the probability of allele 2 at marker *l*. A maximum likelihood estimate of x_{11} was obtained using an expectation maximum approach reviewed by Foulkes (2009). All calculations of the r^2 statistic were implemented in the R package *genetics* (Warnes et al., 2011).

Missing data simulation

For each of the WW, SW, DTM, and NAB datasets three versions of the genotypic data, summarized in Table 2.1, were created with different levels of simulated missing data. In each of the versions: NA20, NA50, and NA70, missing values were introduced at random but the maximum percent missing data at a given marker was set to 20%, 50% and 70% respectively. Examples of the simulated markers sets are illustrated in supplemental Figure 2.1 (Figure S2.1). A total of 10 replicates of each simulated dataset were created and the mean percent of total data points that are missing across the 10 replicates is shown in Table 2.1. The distribution of per-marker percent missing values from the SRRW data versions NA20, NA50, NA70 were used to assign the percent missing at each marker for each of the WW, SW, DTM, and NAB datasets to produce versions NA20, NA50 and NA70, respectively. Across all the missing data versions of all the datasets, the percent missing per marker distribution had a long left tail and a large concentration of values near the threshold level.

Imputation methods

In all cases, the genotypic data were considered continuous variables. The methods MNI, kNNI, SVDI, EMI, and RFI were used to impute the simulated missing values. For all methods the input was an *m* x *n* genotype matrix **M** with *m* individuals and *n* markers. For MNI, each missing data-point x_{ij} at a given marker *j* was replaced with the mean of the non-missing values at that marker.

For kNNI (Troyanskaya et al., 2001), the data points were imputed by replacing them with the weighted average of the data points at the k closest markers. Euclidean distance was used as the measure of marker distance. Euclidean distance between marker genotype vectors **q** and **v** of length *m* was defined as*:*

$$d(\mathbf{q},\mathbf{v}) = \sqrt{(\mathbf{q}_1 - \mathbf{v}_1)^2 + (\mathbf{q}_2 - \mathbf{v}_2)^2 + ... + (\mathbf{q}_m - \mathbf{v}_m)^2}$$

In detail, 1) missing values were first replaced using MNI and the Euclidean distance between all of possible pairs of marker vectors was computed. Each marker was included in the marker matrix twice, both in its original and flipped state to ensure that markers in negative LD would not be considered distant to the marker of interest. 2) For each marker *j*, markers were sorted based on Euclidean distance to marker *j*. 3) For each row *i* of marker *j* the weighted average of the k closest markers with non-missing values at row *i* were used as an estimate of marker data point x_{ij}. The weight of each marker was $1/d^2$ were *d* is the Euclidean distance between marker *j* and the marker to be weighted. kNNI makes no assumptions about the distribution of the data.

For SVDI (Troyanskaya et al., 2001), a singular value decomposition of genotype matrix **M** was used to obtain a set of the k most significant Eigen-vectors of the markers. These k Eigen-vectors were then used as the predictors for linear regression estimation of the missing data points. SVDI was implemented in R (R Development Core Team, 2011) using the package 'bcv' (Perry, 2009). The genotype matrix M can be described as:

$$\mathbf{M}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}}$$

Where **U** has dimensions *m* x k, **V** has dimensions *n* x k, and **Σ** is a k x k diagonal matrix. **U** contains the left singular vectors with are equivalent to the Eigen-vectors of the markers. The corresponding singular values are in the diagonal elements of **Σ**. The singular values are equivalent to the square root of the Eigen-values. The k most significant Eigen-vectors of the markers were those with the k largest Eigen-values. The imputation procedure is described as follows: 1) Missing values were originally imputed using MNI. 2) Singular value decomposition was used to estimate the k most significant Eigen-vectors of the markers: $\hat{\mathbf{U}}$. 3) For each marker *j*, linear regression coefficients of each column of $\hat{\mathbf{U}}$ were estimated by the multiple linear regression equation:

$$\mathbf{Y} = \hat{\mathbf{U}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Where **Y** is a column vector for marker *j*, $\hat{\mathbf{U}}$ is an *m* x *k* matrix of k Eigen-vectors, $\boldsymbol{\beta}$ is a vector of regression coefficients and **ε** is a random error term. Only individuals with non-missing values in Y were used to estimate **β**. 4) $\hat{\mathbf{U}}$ and the

estimates of the regression coefficients, $\hat{\boldsymbol{\beta}}$, were used to estimate the missing values at marker *j*. 5) Using the current version of the genotype matrix the steps two through four were repeated for a total of 10 iterations which was sufficient to meet the convergence criteria which was:

$$\frac{\left|RSS_0 - RSS_1\right|}{RSS_1} < 0.02$$

RSS is the residual sum of squares between the non-missing values and their SVDI model approximation. RSS_0 and RSS_1 are the *RSS* values of successive iterations. SVDI assumes that the genotype matrix is multivariate normal distributed. For the optimal k value calculation methods and results for both kNNI and SVDI see the supplemental information. Optimal k values are listed in supplemental table 12. (Table S2.1).

For EMI, the non-missing marker data was used to obtain maximum likelihood estimates of the vector of means, $\hat{\mathbf{u}}$, and covariance matrix $\hat{\mathbf{X}}$ of the individuals based on the markers. These estimates were then used to obtain multiple linear regression estimates of the missing marker values. $\hat{\mathbf{u}}$, and $\hat{\mathbf{X}}$ were then re-estimated and were used to re-estimate the missing marker values. This process was repeated until the difference between the new estimate and the previous estimate of $\hat{\mathbf{u}} + \hat{\mathbf{X}}\hat{\mathbf{X}}^{\mathrm{T}}$ was 0.02 or less. EMI was implemented using the R package *rrBLUP* (Endelman, 2011). For a more detailed description of this EMI algorithm refer to Poland et al. (2012). For a more through description of the EM imputation algorithm in general refer to Dempster et al. (1977).

For RFI, missing marker values were estimated using random forest regression (Breiman, 2001) using all available data to predict the missing values for every marker. RFI was implemented in R (R Development Core Team, 2010) using the package *MissForest* (Stekhoven and Bühlmann, 2011). The imputation procedure was: 1) for marker matrix M, markers were sorted from lowest to highest percent missing and missing values were imputed using MNI. 2) At each marker *j* containing missing values, the non-missing values, **Y**, were used to grow 100 random forest regression trees Θ1… Θ100. Each tree was grown using a bootstrapped sample of individuals Y and a random sample of $\sqrt{n-1}$ marker predictors were used where *n*-1 is the number of markers excluding marker *j*. Each tree Θ contains the terminal node values and a set of instructions for recursively partitioning the observations into the terminal nodes: these instructions include the split variables at each node, and the value of the split variable used for partitioning. 3) Missing values at marker *j* were imputed as:

$$\hat{Y} = \frac{1}{100}\sum_{1}^{100} h(x,\Theta)$$

where *x* are the input variables.

4) Marker *j* was then updated in marker matrix **M** by using the $\hat{Y}$ values as the estimate of the missing values. 5) Step two through four were repeated for each subsequent marker until all markers were imputed. 6) Then, using this imputed matrix, steps two through five were repeated until convergence or for a maximum of ten iterations. Convergence was declared as soon as the ΔN

increased for the first time where:

$$\Delta N = \frac{\sum_{j \in n} (\mathbf{M_1} - \mathbf{M_0})^2}{\sum_{j \in n} (\mathbf{M_1})^2}$$

$\mathbf{M_1}$ and $\mathbf{M_0}$ are the newly imputed and previously imputed marker matrices respectively. If the convergence criterion was met, $\mathbf{M_0}$ was used as the final estimate of **M**. RFI makes no assumptions about the distribution of the data.

Imputation accuracy calculations

The per-marker imputation accuracy, R_m^2, was described using the R^2 value between predicted data points and the original data points for a given marker vector or individual vector x of length j. The R^2 was defined as

$$R^2 = 1 - \frac{\sum_j (x_j \text{true} - x_j \text{imputed})^2}{\sum_j (x_j \text{true-mean}(x))^2}$$

The R_m^2, as well as the imputation R^2 of the individual genotypes, referred to as R_i^2, were calculated. For each dataset and missing data level, average R_i^2 and R_m^2 across the 10 missing data simulations were also calculated and referred to as $\overline{R}_i^2$ and $\overline{R}_m^2$.

In order to compare with imputation accuracies reported in other publications, for each $\overline{R}_m^2$ value, the equivalent percent correct was also calculated. Because imputed values were continuous, the percent correct for each marker could not be directly calculated. Instead, for each marker, equivalent percent correct values were determined by simulation using each marker's MAF

and $\overline{R}_m^2$ supplemental information.

Computational time

For the first replicate of simulated missing datasets, whenever a dataset was imputed, the number of seconds required for imputation to be completed using one central processing unit (CPU) was recorded. All jobs were submitted to the Computational Biology Service Unit at Cornell University which uses 1) a 240 core Windows cluster consisting of 60 Dell PowerEdge 1855 nodes with two x64 Pentium 4 Xeon 3.4GHz, 4GB RAM and 144GB HD each and 2) a 400 core Windows cluster consisting of 200 Sun V20Z nodes with two AMD Opteron 248 2.2GHz, 2GB RAM and 300GB HD each.

Assessment of factors affecting imputation accuracy

For each imputation method factors affecting the imputation accuracy were assessed. A marker's minor allele frequency (MAF), number of non-missing data points, and level of LD with other markers were considered as factors that could impact its imputation accuracy. The distance between an individual and its closest relative and the expected prediction error variance (PEV) were considered as factors affecting the imputation accuracy on an individual genotype basis. The impact of each of these factors was assessed for each imputation method using the WW, SW, DTM, and NAB datasets post imputation.

First, the impact of MAF on the imputation accuracy was assessed. For each dataset-imputation method combination, $\overline{R}_m^2$ was averaged across dataset

versions NA20, NA50 and NA70 and this overall estimate of marker imputation accuracy is referred to as $\overline{\overline{R}}_m^2$.The median $\overline{\overline{R}}_m^2$ for each value of MAF rounded to the nearest tenth was calculated. The relationship between the median $\overline{\overline{R}}_m^2$ and the MAF value was then plotted to characterize the relationship.

The impact of the number of non-missing data points at a marker on the marker's imputation accuracy was assessed for each dataset-imputation method combination using data from all 10 replicates and versions NA20, NA50 and NA70 combined. For each marker, the number of non-missing data points was rounded to the nearest factor of 5, and for each value the median R_m^2 was calculated.

To determine the impact of the LD level with other markers on the imputation accuracy, markers were first classified as markers in low LD with all other markers or markers in at least moderate LD with at least one other marker. Markers whose highest r^2 statistic was less than 0.5 were considered to be in low LD with all other markers. A marker that had at least one r^2 statistic greater than or equal to 0.5 was considered to be in at least moderate LD with at least one other marker. The median $\overline{\overline{R}}_m^2$ of markers in low LD and of markers in at least moderate LD with at least one other marker was calculated. The ratio of $\overline{\overline{R}}_m^2$ for markers in low LD to the $\overline{\overline{R}}_m^2$ for markers in at least moderate LD was then examined.

To assess the effect of the genetic distance between an individual and its

closest relative on the individual genotype imputation accuracy, the Euclidian distance was calculated for each pair of individuals and the R_i^2 of each dataset was measured for each simulated dataset and imputation method combination. The mean R_i^2 values across all replicates, $\overline{R}_i^2$, were averaged across versions NA20, NA50, and NA70 of a given dataset-imputation method combination to calculate an overall mean R_i^2 for each individual which is referred to as $\overline{\overline{R}}_i^2$. The Euclidian distance between each individual and its closest relative, rounded to the nearest whole number was plotted against the median $\overline{\overline{R}}_i^2$ to examine the relationship.

The relationship between PEV for the genetic values and the $\overline{\overline{R}}_i^2$ was also examined. An individual's PEV is a measure of genetic connectedness to the other individuals (Kennedy and Trus, 1993) where an individual's connectedness is determined by the number and strength of the genetic relationships between that individual and the other individuals in the dataset. For example, a low PEV indicates high connectedness and high degree of genetic relationship. To measure an overall PEV value for each individual, a vector of PEVs was calculated for each marker using the mixed model equations (Searle et al. 1992) implemented in the R package *rrBLUP* (Endelman, 2011). The genetic and error variance components were estimated using maximum likelihood and the genomic relationship matrix, excluding the response variable marker, was used as the covariance matrix between genotypes. The sum of the PEV vectors across all markers was used as

the overall PEV vector. Because PEV is a reflection of the number and strength of the genetic relationships between individuals, it is expected to be a useful indicator for how well and individual's missing data can be imputed using all other individuals as a reference.

Genomic Selection accuracy calculation

All 10 simulations of missing data versions NA20, NA50, and NA70 of the WW, SW, DTM, NAB and SRRW marker sets were imputed with each of the imputation methods: MNI, kNNI, SVDI, EMI, and RFI. Then, each of the 10 replicates of the marker set-imputation method combinations was used to calculate the 10-fold cross validation GS accuracy for both ridge regression (Whittaker et al., 2000) and Bayesian Lasso (de los Campos et al., 2009), see supplemental information. GS accuracies are computed as the Pearson's correlation between the phenotype estimated breeding values and the GEBVs. The mean accuracy for each marker set-imputation method-prediction model combination was computed. GS accuracies were also computed using version NA0 of the WW, SW, DTM, and NAB genotypic data.

Results

Linkage disequilibrium between markers

For each dataset, the LD between marker pairs was quantified using the r^2 statistic. Markers that had at least one other marker associated with $r^2 \geq 0.5$ were considered to be in at least moderate LD with at least one other marker. In the

WW, SW, DTM, and NAB datasets, 62%, 74%, 12% and 69% of the markers had at least one other marker in at least moderate LD, respectively. Comparatively, LD between markers was high in the SW, NAB, and WW datasets and much lower in the DTM dataset.

Imputation accuracy

The imputation accuracy reported as the median $\overline{R}_m^2$ is shown in Figure 2.1 for kNNI, SVDI, RFI, and EMI.

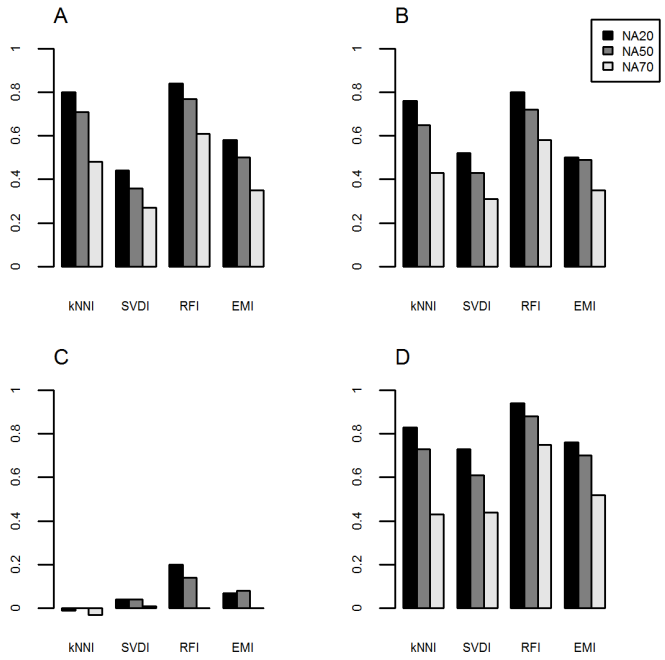


Figure 2.1: Median $\overline{R}_m^2$ of each imputation method across all datasets

(A) Cornell winter wheat (WW), (B) CIMMYT elite spring wheat (SW), (C) CIMMYT drought tolerant maize (DTM), (D) North American barley (NAB). For each population median R_m^2 obtained using k-nearest neighbors imputation (kNNI), singular value decomposition imputation (SVDI), random forest regression imputation (RFI), and expectation maximization imputation (EMI), are shown for the three dataset versions: NA20 (black), NA50 (grey), and NA70 (white) which contain up to 20%, 50%, and 70% missing values per marker, respectively.

For all dataset-imputation method combinations, $\overline{R}_m^2$ values were non-normal and there were many extreme values. The median $\overline{R}_m^2$ values and the equivalent percent correct values are listed in Table 2.2.

Table 2.2: Median $\overline{R}_m^2$ and median percent correct† for each imputation method and across all datasets

		Imputation Method‡			
Dataset§	Version¶	kNNI	SVDI	RFI	EMI
WW	NA20	0.8 / 97	0.44 / 93	0.84 / 98	0.58 / 95
	NA50	0.71 / 96	0.36 / 92	0.77 / 97	0.5 / 93
	NA70	0.48 / 94	0.27 / 89	0.61 / 95	0.35 / 91
	Mean	0.66 / 96	0.36 / 91	0.74 / 97	0.48 / 93
SW	NA20	0.76 / 96	0.52 / 93	0.8 / 97	0.5 / 93
	NA50	0.65 / 95	0.43 / 93	0.72 / 96	0.49 / 93
	NA70	0.43 / 93	0.31 / 91	0.58 / 94	0.35 / 91
	Mean	0.61 / 95	0.42 / 92	0.7 / 96	0.45 / 92
DTM	NA20	-0.01 / 82	0.04 / 83	0.2 / 88	0.07 / 85
	NA50	0 / 82	0.04 / 83	0.14 / 87	0.08 / 84
	NA70	-0.03 / 82	0.01 / 83	0 / 84	0 / 83
	Mean	-0.01 / 82	0.03 / 83	0.11 / 86	0.05 / 84
NAB	NA20	0.83 / 99	0.73 / 98	0.94 / 100	0.76 / 98
	NA50	0.73 / 99	0.61 / 98	0.88 / 99	0.7 / 98
	NA70	0.43 / 97	0.44 / 97	0.75 / 99	0.52 / 97
	Mean	0.66 / 98	0.59 / 98	0.85 / 99	0.66 / 98

†Median $\overline{R}_m^2$ and median percent correct are separated by a backslash (/)

‡kNNI: k-nearest neighbors imputation, SVDI: singular value decomposition imputation, EMI: expectation maximization imputation, RFI: random forest regression imputation

§WW: Cornell winter wheat, SW: CIMMYT elite spring wheat, DTM: CIMMYT drought tolerant maize, NAB: North American barley, SRRW: CIMMYT stem rust resistant wheat

¶NA20: up to 20% missing data per marker, NA50: up to 50% missing data per marker, NA70: up to 70% missing data per marker

The population with the highest median $\overline{R}_m^2$ for each of the levels of missing data was the NAB population, while the lowest imputation accuracies were observed

with the DTM population. As expected, median $\overline{R}_m^2$ values always decreased as the level of missing data increased. RFI always produced the highest accuracies; kNNI generally produced the second highest accuracies, followed by EMI and SVDI. The rankings were slightly different for the DTM dataset, where RFI was most accurate followed by EMI, SVDI, and kNNI. The rankings of the methods for each dataset according to the median percent correct are the same as those according to the median $\overline{R}_m^2$, however the median percent correct values could not be compared across datasets because percent correct values are influenced by the MAF which differs among datasets.

Computational time

Large differences in the computational requirements for the imputation methods were observed (Table 2.3). kNNI, SVDI, and EMI required relatively little computation time on average, while RFI required at 95x, 760x, and 65x more computation time than kNNI, SVDI, and EMI respectively. For SVDI and kNNI, the computation time required for determining optimal k values was not included in the estimates of the average computational time because the computation time for optimal k estimation depends on the method used for estimation. The 10-fold cross validation approach that we used to estimate optimal k values for SVDI and kNNI requires approximately 50 runs of the SVDI and kNNI respectively. If the time required to estimate optimal k values for SVDI and kNNI were included in the total computational time, EMI would be the fastest of the four imputation methods.

Table 2.3: CPU† minutes required to complete the imputation of one dataset

		Imputation Method‡			
Dataset§	Version¶	kNNI	SVDI	RFI	EMI
WW	NA20	2.5	0.4	364.8	2.2
	NA50	4.7	0.4	411.6	3.1
	NA70	5.6	0.4	280.2	2.7
SW	NA20	5.3	1.5	132.6	5.5
	NA50	9.7	1.5	935.4	9.1
	NA70	11.5	1.5	610.2	7.3
DTM	NA20	1.7	0.2	271.8	0.8
	NA50	3.3	0.2	440.4	0.8
	NA70	4.1	0.2	223.8	1.0
NAB	NA20	24.4	6.0	4084.8	64.6
	NA50	45.1	5.8	4204.2	106.7
	NA70	50.3	5.8	2349	86.2
SRRW	NA20	7.1	0.7	2364.6	3.5
	NA50	14.2	0.6	1618.8	4.8
	NA70	17.1	0.6	1309.2	4.1

† CPU: central processing unit
‡ kNNI: k-nearest neighbors imputation, SVDI: singular value decomposition imputation, EMI: expectation maximization imputation, RFI: random forest regression imputation
§WW: Cornell winter wheat, SW: CIMMYT elite spring wheat, DTM: CIMMYT drought tolerant maize, NAB: North American barley, SRRW: CIMMYT stem rust resistant wheat
¶NA20: up to 20% missing data per marker, NA50: up to 50% missing data per marker, NA70: up to 70% missing data per marker.

Factors affecting imputation accuracy

MAF: For all datasets, $\overline{\overline{R}}_m^2$ values for markers with MAF<0.1 were low compared to that of markers with MAF>0.1; however, the relationship between MAF and $\overline{\overline{R}}_m^2$ for markers with MAF>0.1 was different for each dataset (Figure 2.2). In general, $\overline{\overline{R}}_m^2$ increased as MAF increased as long as MAF<0.4; however, with the NAB dataset (Figure 2.2, D) there was no relationship between MAF and $\overline{\overline{R}}_m^2$ for MAF>0.1. Accuracy in terms of percent correct had a strong negative

linear relationship with the MAF across all imputation methods and datasets. Markers with lower MAF values tended to have higher percent correct values (data not shown).

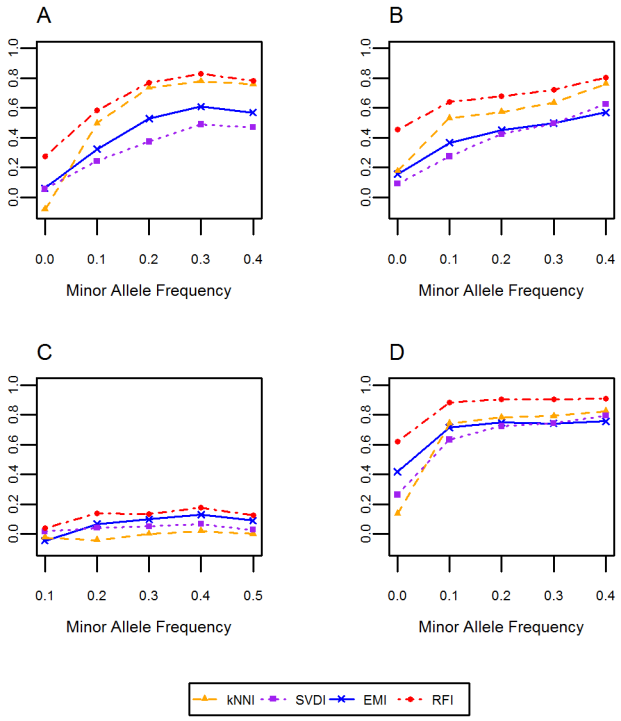


Figure 2.2: Relationship between the minor allele frequency (MAF) and $\overline{\overline{R}}_m^2$

The median $\overline{\overline{R}}_m^2$ obtained for a given MAF rounded to the nearest tenth is plotted for each dataset: (A) Cornell winter wheat (WW), (B) CIMMYT elite spring wheat (SW), (C) CIMMYT drought tolerant maize (DTM), (D) North American barley (NAB). Each color and symbol represents a different imputation method: k-nearest neighbors imputation (kNNI, orange triangles), singular value decomposition imputation (SVDI, purple squares), random forest regression imputation (RFI, red circles), and expectation maximization imputation (EMI, blue crosses).

Number of non-missing data points: With almost all dataset-imputation method combinations, as the number of non-missing data points increased, the, the $\overline{R}_m^2$ levels increased in a linear fashion (Figure 2.3). The strength of this linear relationship was similar for all imputation methods; however, with the DTM

dataset, $\bar{R}_m^2$ for kNNI and SVDI were close to zero regardless of the number of non-missing data points.

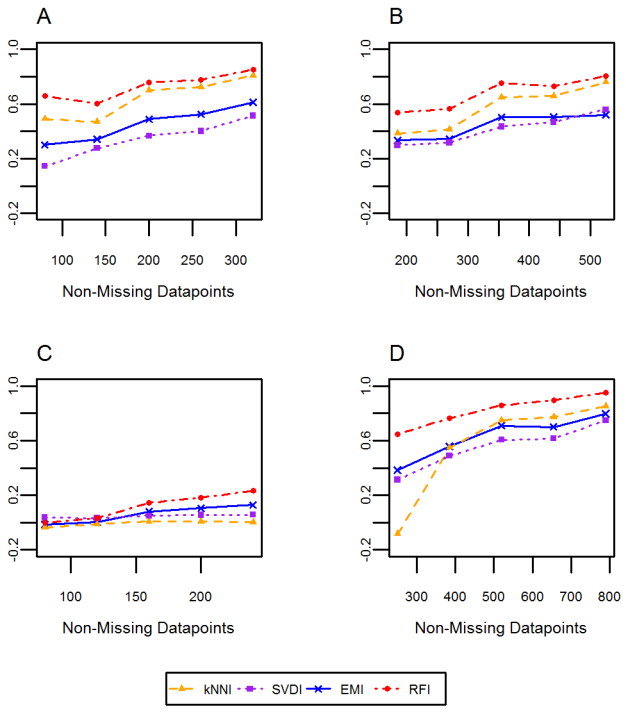


Figure 2.3: Relationship between the number of non-missing data- points and R_m^2

The median R_m^2 obtained for a given number non-missing data points rounded to the nearest factor of 5, is plotted for each dataset: (A) Cornell winter wheat (WW), (B) CIMMYT elite spring wheat (SW), (C) CIMMYT drought tolerant maize (DTM), (D) North American barley (NAB). Each color and symbol represents a different imputation method: k-nearest neighbors imputation (kNNI, orange triangles), singular value decomposition imputation (SVDI, purple squares), random forest regression imputation (RFI, red circles), and expectation maximization imputation (EMI, blue crosses).

LD between markers: The ratio of the median imputation $\bar{\bar{R}}_m^2$ for markers with no other markers in moderate LD to the median imputation $\bar{\bar{R}}_m^2$ for markers with at least one other marker in moderate LD was always less than one (Table 2.4), indicating that the imputation accuracy for markers without markers

in moderate LD was always lower than that for markers that had at least one other marker in moderate LD.

Table 2.4: Ratios† of median $\bar{\bar{R}}_m^2$ of markers having no markers in moderate linkage disequilibrium (LD)‡ to that of markers with at least one other marker in moderate LD

†Reduced ratios are reported followed by the values used to compute the reduced ratios in parenthesis
‡at least moderate LD was defined as r^2 statistic ≥ 0.5

	Imputation Method§			
Dataset¶	kNNI	SVDI	RFI	EMI
WW	0.16 (0.13/.8)	0.36 (0.17/0.47)	0.49 (0.41/0.84)	0.39 (0.23/0.59)
SW	0.14 (0.1/0.7)	0.47 (0.23/0.49)	0.62 (0.47/0.76)	0.58 (0.29/0.5)
DTM	-0.18 (-0.03/0.17)	0.33 (0.02/0.06)	0.18 (0.09/0.5)	0.14 (0.03/0.22)
NAB	0.31 (0.24/0.78)	0.59 (0.40/0.68)	0.74 (0.67/0.9)	0.63 (0.46/0.73)
Mean	0.11	0.44	0.51	0.44

§kNNI: k-nearest neighbors imputation, SVDI: singular value decomposition imputation, EMI: expectation maximization imputation, RFI: random forest regression imputation
¶WW: Cornell winter wheat, SW: CIMMYT elite spring wheat, DTM: CIMMYT drought tolerant maize, NAB: North American barley

Across all datasets, the $\bar{\bar{R}}_m^2$ ratios for the two classes of markers was much smaller for kNNI compared to the other imputation methods, indicating that the imputation accuracy of kNNI was more strongly influenced by the level of LD between markers compared to the other methods. With the WW, SW, and NAB datasets the $\bar{\bar{R}}_m^2$ ratios for the two classes of markers was similar for SVDI, RFI, and EMI indicating that the accuracy of these three methods is influenced by the level of LD between markers to a similar degree. However, with the DTM dataset, the $\bar{\bar{R}}_m^2$ ratio for the two classes of markers was closer to one for SVDI compared

to the other methods, indicating that for this dataset, the accuracy with SVDI was less affected by the LD between markers, compared to the other methods.

Distance from the closest relative and PEV: Regardless of the dataset or the imputation method, the smaller the distance between an individual and its closest relative, the higher the $\bar{\bar{R}}_i^2$ (Figure 2.4). One exception was observed with the DTM dataset, where for kNNI there was no relationship between the distance between an individual and its closest relative and $\bar{\bar{R}}_i^2$. We observed very similar trends between $\bar{\bar{R}}_i^2$ and the overall PEV (Figure S2.2). As an individual's PEV increased, indicating a decrease in the strength and number of genetic relationships between that individual and all other individuals, its $\bar{\bar{R}}_i^2$ decreased in all cases except when the DTM dataset was imputed with kNNI.

Effect of imputation method on genomic selection accuracy

In nearly all cases, GS accuracies did not differ greatly from one imputation method to another, with the exception of MNI, which sometimes led to much lower accuracies compared to all other methods when the NA70 dataset version was used (Figure 2.5 and 2.6). Overall, GS accuracies were least affected by the imputation method for dataset version NA20, and most affected by the imputation method for dataset version NA70. The relative performance of each method in terms of GS accuracy after imputation depended on the dataset, and dataset version; however, RFI consistently performed well across all datasets. For

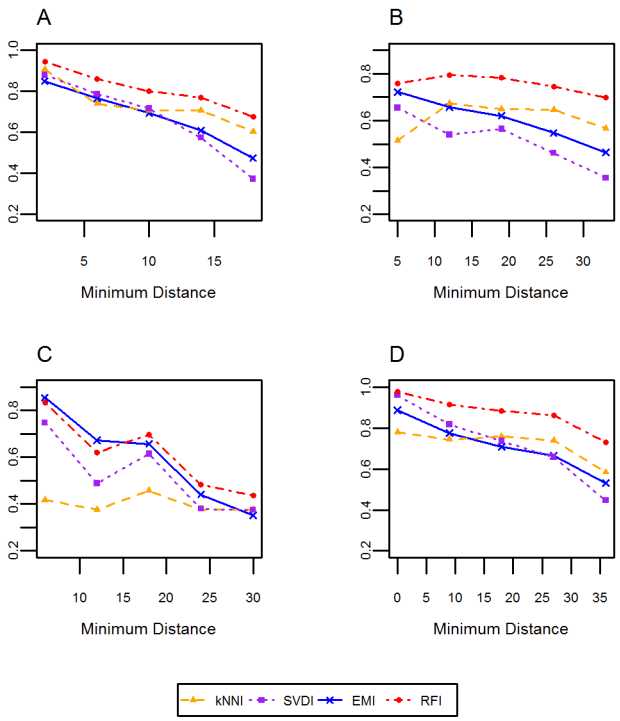


Figure 2.4: Relationship between the distance from the closest relative and $\bar{\bar{R}}_i^2$

The median $\bar{\bar{R}}_i^2$ obtained for a given Euclidean distance between an individual and its closest relative rounded to the nearest whole number is plotted for each dataset: (A) Cornell winter wheat (WW), (B) CIMMYT elite spring wheat (SW), (C) CIMMYT drought tolerant maize (DTM), (D) North American barley (NAB). Each color and symbol represents a different imputation method: k-nearest neighbors imputation (kNNI, orange triangles), singular value decomposition imputation (SVDI, purple squares), random forest regression imputation (RFI, red circles), and expectation maximization imputation (EMI, blue crosses).

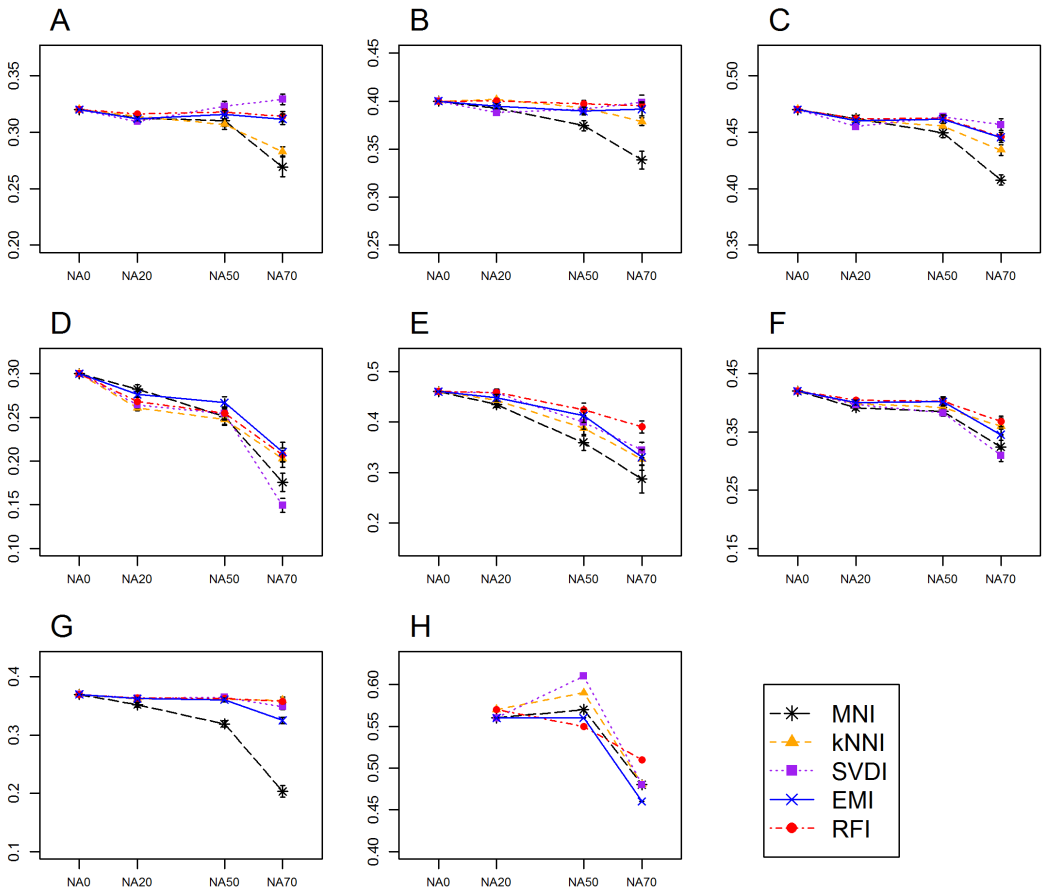


Figure 2.5: Genomic selection (GS) accuracy obtained using ridge regression after imputation

Mean GS accuracies obtained using the dataset versions NA0, NA20, NA50, having up to 0%, 20%, 50%, and 70% missing data per marker respectively, imputed with either mean imputation (MNI; black stars), k-nearest neighbors imputation (kNNI; orange triangles), singular value decomposition imputation (SVDI; purple squares), expectation maximization imputation (EMI; blue crosses) and random forest regression imputation (RFI; red circles) are shown for (A) Cornell winter wheat (WW)-yield, (B) Cornell winter wheat (WW)-height, (C) Cornell winter wheat (WW)-protein, (D) Cornell winter wheat (WW)-days to heading (DTH), (E) CIMMYT drought tolerant maize (DTM), (F) CIMMYT elite spring wheat (SW), (G) North American barley (NAB), and (H) stem rust resistant wheat (SRRW) datasets. Each plot has a different y-axis range. Error bars depict standard errors.

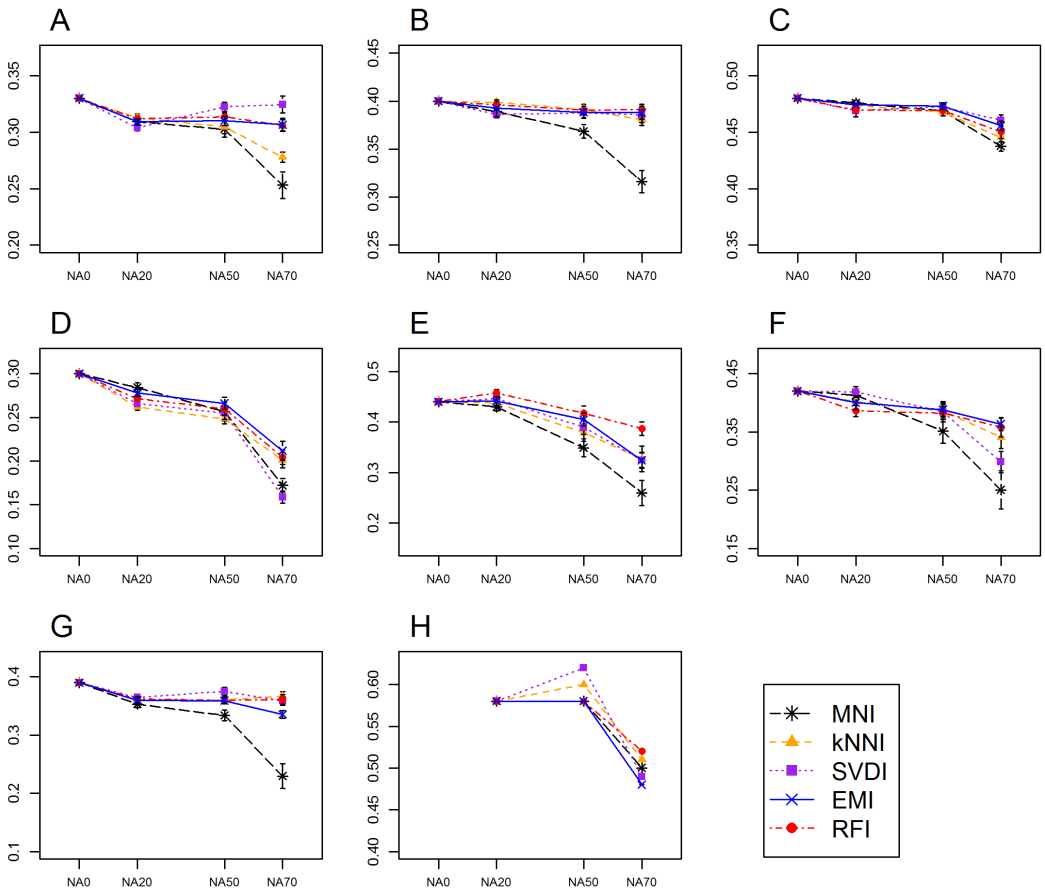


Figure 2.6: Genomic selection (GS) accuracy obtained using Bayesian Lasso after imputation

Mean GS accuracies obtained using the dataset versions NA0, NA20, NA50, having up to 0%, 20%, 50%, and 70% missing data per marker respectively, imputed with either mean imputation (MNI; black stars), k-nearest neighbors imputation (kNNI; orange triangles), singular value decomposition imputation (SVDI; purple squares), expectation maximization imputation (EMI; blue crosses) and random forest regression imputation (RFI; red circles) are shown for (A) Cornell winter wheat (WW)-yield, (B) Cornell winter wheat (WW)-height, (C) Cornell winter wheat (WW)-protein, (D) Cornell winter wheat (WW)-days to heading (DTH), (E) CIMMYT drought tolerant maize (DTM), (F) CIMMYT elite spring wheat (SW), (G) North American barley (NAB), and (H) stem rust resistant wheat (SRRW) datasets. Each plot has a different y-axis range. Error bars depict standard errors.

the WW datasets, the relative performance of the imputation methods in terms of GS accuracy was inconsistent across the four traits tested. For, a given dataset and dataset version, the rank of each method based on $\overline{R}^2_m$, was not consistent with the rank based on GS accuracy using RR or BL post-imputation. The rank of the imputation methods, however, was consistent between the two different GS models. We also found that including rather than removing 'sparse' markers, those with large amounts of missing data, nearly always led to higher GS accuracies (methods and results described in supplemental information), especially when RFI, kNNI or EMI were the imputation methods used (Figure S2.4).

Discussion

Imputation accuracy

This study found that map-independent imputation methods other than MNI can be surprisingly accurate, especially when LD between markers is high and the genotyped individuals are related. RFI was the most promising method overall because of its consistently high performance in terms of imputation accuracy and subsequent GS accuracy; however, it was the most computationally intensive method evaluated. kNNI, while less accurate than RFI, may be a good alternative to RFI if there are computational limitations to completing the imputation. It is likely that RFI and kNNI produced comparable levels of accuracy because both use a similar model free approach for imputation

that involves weighting a selected set of k important variables according to a distance metric (Lin and Jeon, 2006). The weighted average of these variables is the predicted value of the variable of interest. For kNNI the distance metric was the Euclidean distance and k was a fixed number across all variables. For RFI, the k important variables and their weights are determined by the splitting scheme of the tree that is determined using the response variable. The increased accuracy but greater computational burden of the RFI method compared with kNNI is due to its adaptive weighting of variables that takes into account the response variable.

A possible reason that EMI and SVDI were less accurate than RFI and kNNI, is that the genotypic datasets that we used may have violated multivariate normality, an underlying assumption for EMI and SVDI. Alternatively, EMI and SVDI may not have been as effective at ignoring uninformative predictors. If true, linear regression based imputation methods involving variable selection could be as accurate as kNNI or RFI. However, due to multicollinearity, attempts to test imputation based on subset selection methods such as stepwise regression were not successful. Regression imputation using variable selection methods which can cope with multicollinearity, such as the least absolute shrinkage and selection operator (Lasso; Tibshirani, 1996), would be interesting to test in future studies.

EMI performed consistently better than SVDI which is likely because EMI incorporates all the marker data as predictors whereas SVDI first used a data reduction step, potentially eliminating useful information. SVDI may have

outperformed EMI if the datasets had a greater rate of genotyping error because it is expected to better cope with noisy data (Troyanskaya et al., 2001).

For all methods, average median imputation accuracies on an individual genotype basis $\bar{\bar{R}}_i^2$ were not always homogenous across population sub-groups as illustrated in Figure S5, which shows individuals plotted according to the first two principal components of their marker genotypes and color coded according to their imputation accuracy. With the DTM and WW datasets, small sub-groups of individuals that clustered together according to the first two principal components of marker genotypes tended to have similar ranges of accuracy. However, with the SW and NAB datasets $\bar{\bar{R}}_i^2$ was relatively homogenous across population sub-groups. An association between $\bar{\bar{R}}_i^2$ and population sub-group is undesirable because it may create or worsen an association between GS accuracy and population sub-group. Using large datasets with minimal population structure for imputation and genomic selection is advocated to avoid heterogeneity of imputation and genomic selection accuracies across sub-groups of individuals.

Population structure may also lead to increased imputation accuracy for markers with high levels of population subdivision (Iwata and Jannink, 2010) because an individual's allelic state can be predicted largely based population sub-group alone. Accuracy levels for datasets with many markers highly subdivided by population may be high largely because of structure; we therefore

calculated $\bar{\bar{R}}_m^2$ excluding markers with high levels of population subdivision as indicated by their Fst values, where high Fst indicates high population sub-division (for methods see supplemental information.) For markers with MAF>0.1, on average, $\bar{\bar{R}}_m^2$ excluding markers with the 25% highest Fst values were 0.9, 1.17, 1.02, and 0.9 times those of overall $\bar{\bar{R}}_m^2$ for the WW, SW, DTM, and NAB datasets respectively. Thus, for the WW and NAB datasets, the high imputation accuracies we observed may have been in small part due to population structure.

Comparing our imputation accuracy results with those of other studies is difficult because each study uses different populations of different sizes, levels of missing data, MAF distributions, and levels of LD between markers. In addition, accuracy reported as percent correct cannot be compared across datasets with different MAF distributions. Nevertheless, we assume that map-dependent imputation methods would outperform the map-independent methods that we evaluated (given the availability of an accurate genetic or physical map) because physically linked markers are used to predict missing values. These physically linked markers should be more reliable predictors compared to markers that are in LD but may not be physically linked. As genetic and physical maps develop for wheat and barley the assumption that map-dependent methods would outperform the map-independent methods can be tested.

Factors affecting imputation accuracy

Markers with very low MAF had low $\bar{\bar{R}}_m^2$ values. There are two possible

explanations for this observation. First, because of the way R_m^2 is calculated, a single imputation error has a much larger negative impact on the R_m^2 for markers with lower MAF values (Figure S6). Thus, it is harder to achieve high R_m^2 for markers with a low MAF. Second, individuals with the minor allele at a given marker are not well represented, making their marker genotype more difficult to predict. A similar relationship between MAF and R_m^2 was also found by studies by Iwata and Jannink (2010) and Li et al., (2011) which used map-dependent imputation methods. Unlike R_m^2, imputation accuracy in terms of percent correct had a negative linear relationship with MAF (data not shown), this is because markers with lower MAF can always be imputed with a reasonably high percent correct based on the marker mean alone. Other studies of map-dependent imputation methods report a negative relationship between MAF and percent correct (Pei et al., 2008; Hickey et al., 2012).

The number of non-missing data points, analogous to reference panel size in other studies, was found to positively impact the R_m^2. This finding is consistent with other studies which tested the effect of reference panel size on the imputation accuracy using map-dependent methods (Pei et al., 2008; Druet et al., 2010; Li et al., 2010). For RFI, EMI, and SVDI, which involve a model training step, fewer missing data points means that more individuals are available for model training. With kNNI, a smaller number of non-missing data points at a given marker leads to a more accurate estimate of its distance from all other markers.

However, with the DTM set there was no trend between accuracy and the number of non-missing data points with kNNI. This was because accuracy with kNNI for this dataset was very low overall.

The presence of one or more markers in moderate LD (r^2 statistic ≥ 0.5) was a more important factor for kNNI compared to RFI, EMI, and SVDI. This is because kNNI bases its predictions on a fixed number of close markers, whereas RFI, EMI and SVDI use information from all markers in the dataset to generate predicted values for the missing data points. The LD between markers on a whole dataset basis also appeared to be an important factor affecting the $\bar{\bar{R}}_m^2$ of all methods because accuracies with the DTM dataset, which had low levels of LD between markers overall, were much lower than accuracies with the WW, SW, and NAB datasets. Other publications that have evaluated the effect of LD on imputation accuracy for map-dependent methods have found similar trends (Pei et al., 2008; Hickey et al., 2012).

We found that imputation accuracy on an individual genotype level was negatively correlated with the distance from the closest relative in the dataset, and the PEV, which is an indication of the relationship between an individual and other genotypes. A similar relationship between imputation accuracy and relationship has been found by other studies of map-dependent imputation methods (Druet et al., 2010; Zhang and Druet 2010; Hickey et al., 2012). It is clear that to ensure effective imputation, the dataset to be imputed should contain related individuals. If the dataset is suited for GS, it is likely that the individuals

are already related. However, to increase the chances that an individual will have close relatives in the dataset, all available genotypic data for the germplasm pool of interest should be combined prior to imputation.

Genomic selection accuracy

The GS accuracies that we observed may be sufficiently high to lead to increased rates of genetic gain compared to phenotypic selection (PS), depending on the accuracy of PS and the selection cycle duration of both PS and GS. It is important to note that all GS accuracies reported for a given dataset are 'global estimates' across all potential sub-populations. Based on other studies evaluating GS accuracies within and across sub-populations (Zhao et al.2011, Heslot et al., 2012, Windhausen et al. in 2012), this global accuracy estimate may be greater than the accuracy measured within individual sub-populations.

Effect of imputation method on the genomic selection accuracy

Improved accuracy of GS after application of map independent imputation methods was another important finding of this study. Based on our results, unordered markers with missing data can be included in the dataset to improve accuracy through imputation with RFI, kNNI, EMI, or even SVDI rather than MNI. However, for datasets with low levels of missing data (up to 20% per marker), imputing with MNI is sufficient. Although our results do not support removing markers with high levels of missing data prior to GS, in many datasets markers with low levels of missing data may be sufficient to saturate the genome. With the datasets used in this study, the average number of markers with up to 20% and

50% missing data was 18 to 37 and 99 to 186 respectively, and these reduced marker sets were not sufficient to saturate the genome. Thus, including markers with larger amounts of missing data led to improved GS accuracies. Interestingly, a low median $\overline{R}_m^2$ was not reflective of the merit of imputation prior to GS. The median $\overline{R}_m^2$ for the datasets with up to 70% missing data per marker were the lowest of all the missing data levels; however we saw the greatest gain in GS accuracy from kNNI, SVDI, EMI, or RFI relative to MNI imputation with this level of missing data. This was especially apparent for the DTM dataset, which had a median $\overline{R}_m^2$ near zero for most methods when there was up to 70% missing data per marker. However, RFI on this dataset produced GS model accuracies 1.3 times greater than those achieved when MNI was used prior to GS. Surprisingly, the most accurate imputation method was not always the method that gave the highest GS accuracy. This may be caused by non-random imputation errors. If some imputation errors are similar for related individuals, these non-random errors may able to capture some genetic relationships in the GS model. The idea that the imputation errors may capture some genetic relationships was suggested by a study by Weigel et al. (2010).

Conclusions

This study has important implications for species that lack a reference genome, complete reference map, and pre-designed high-throughput genotyping platforms. First, unordered markers can be imputed with high levels of accuracy,

and even higher accuracies may result if additional reference genotypes can be added to the dataset prior to imputation. Based on the results of this study, if a large number of marker genotypes are produced (so that markers are in LD with each other), and the population contains individuals with some genetic relationship, missing data can be imputed with reasonable accuracy even if the level of missing data is high; up to 70%. Future work to improve upon and reduce the computational burden of the most promising methods in this study, RFI and kNNI, would be especially useful if these methods are to be used widely. The second implication of this study is that a large proportion of missing data in dense marker sets is not a major concern for GS. As long as the marker density is sufficiently high, the accuracy does not appear to be strongly negatively affected. In cases where missing data does negatively impact the GS accuracy imputation using a method other than MNI prior to GS model training and validation can help improve the accuracy. Overall, map-independent imputation shows promise for the feasibility of applying GS, enabled by emergent sequence-based genotyping technologies, to almost any species regardless of the availability of pre-existing genotyping resources.

Acknowledgements

This research was funded by The Bill & Melinda Gates Foundation (Durable Rust Resistance in Wheat) and the United States Department of Agriculture[1]-Agricultural Research Service (USDA-ARS) (Appropriation No. 5430-21000-006-00D).

References

Akbari, M., P. Wenzl, V. Caig, J. Carling, L. Xia, Y. Shiying, U. Grzegorz, M. Volker, L. Anke, K. Haydn, M.J Hayden, N. Hayden, N. Howes, P. Sharp. 2006. Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. Theor. Appl. Genet. 113: 1409-20.

Barley Coordinated Agriculture Project, 2011 Introduction to project. Available at http://www.barleycap.org/ (verified 30 July 2012). Univ. of Minnesota, St. Paul, MN.

Breiman, L., 2001. Random forests. Mach. Learn. 45: 5-32.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, J.M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182: 375-85.

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueño, J. L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, Y. Jianbing. V. Arief, M. Banziger, H.-J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713-724.

Dassonneville, R., R. F. Brøndum, T. Druet, S. Fritz, F. Guillaume, B. Guldbrandtsen, M.S. Lund, V. Ducrocq, G. Su. 2011 Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. J. Dairy Sci. 94: 3679–3686.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B. Met. 39: 1-38.

Druet T., C., Schrooten, and A. P. W., de Roos, 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. J. Dairy Sci. 93: 5443–5454.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E.S. Buckler, S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: e19379.

Endelman, J. B., 2011 Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Gen. 4: 250-255.

Foulkes, A.S. 2009. Applied statistical genetics with R: for population-based association studies. Springer, Berlin.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2009. Genomic selection using low-density marker panels. Genetics 182: 343-53.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard, 2009. Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92: 433-443.

Heffner, E. L., J.-L. Jannink, and M. E. Sorrells, 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Gen. 4: 65-75.

Heffner, E. L., A. J. Lorenz, J.-L. Jannink, and M. E. Sorrells, 2010. Plant breeding with genomic selection: gain per unit time and cost. Crop Sci. 50: 1681-1690.

Heffner, E. L., M. E. Sorrells, and J.-L. Jannink, 2009. Genomic selection for crop improvement. Crop Sci. 49: 1-12.

Heslot, N., H.-P. Yang, M.E. Sorrells, and J.-L. Jannink, 2012. Genomic selection in plant breeding: a comparison of models. Crop Sci. 52: 146–160.

Hickey, J. M., J. Crossa, R. Babu, and G. de los Campos, 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52: 654-663.

Iwata, H., and J.-L. Jannink, 2010. Marker genotype imputation in a low-marker-density panel with a high-marker-density reference panel: accuracy evaluation in barley breeding lines. Crop Sci. 50: 1269-1278.

Kennedy B. W. and D. Trus, 1993. Considerations on genetic connectedness between management units under an animal model. J. Anim. Sci. 71: 2341-2352

Li, L., Y. Li, S. R. Browning, B. L. Browning, A. J. Slater, X. Kong, J. L. Aponte, V .E. Mooser, S. L. Chissoe, J. C. Whittaker, M.R. Nelson, M. G. Ehm. 2011. Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. PLoS One 6: e24945.

Li, Y., C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis, 2010 MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 834: 816-834.

Lin, Y., and Y. Jeon, 2006 Random forests and adaptive nearest neighbors. J. Am. Statist. Assoc. 101: 578-590.

Lorenz, A. J., S. Chao, F. G. Asoro, E. L. Heffner, T. Hayashi, I. Hiroyoshi, K.P. Smith, M.E. Sorrells, J.-L. Jannink. 2011. Genomic selection in plant breeding: knowledge and prospects. Adv. Agron. 110: 77-123.

Lorenzana, R. E., and R. Bernardo, 2009 Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. 120: 151-161.

Marchini, J., and B. Howie, 2010 Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 11: 499-511.

Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly, 2007 A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39: 906-13.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Mulder H. A., M. P. L. Calus, T. Druet, C. Schrooten, 2012 Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. J. Dairy Sci. 95: 876–889.

Pei, Y.-F., J. Li, L. Zhang, C. J. Papasian, and H.-W. Deng, 2008 Analyses and comparison of accuracy of different genotype imputation methods. PloS One 3: e3551.

Perry, P. O., 2009 bcv: Cross-Validation for the SVD. R package version 1.0. Avaliable at: http://CRAN.R-project.org/package=bcv/ (verified 30 July 2012).

Poland, J., and T.W. Rife. 2012 Genotyping-by-sequencing for plant breeding and genetics. Plant Gen. 5: 92-102

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sánchez-Villeda, M. Sorrells, and J.-L. Jannink. 2012. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Gen. 5: 103–113.

R Development Core Team, 2011 R: A Language and Environment for Statistical Computing, Vienna. Available at: http://www.r-project.org/ (verified 30 July 2012).

Searle, S. R., G. Casella and C.E. McCulloch. 1992. *Variance Components*. John Wiley & Sons, Inc., New York

Stekhoven, D. J., and P. Bühlmann, 2011 MissForest - nonparametric missing value imputation for mixed-type data. Bioinformatics (Oxford, England) 28: 112-118.

Tibshirani, R., 1996 Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B. Met**.** 58: 267-288.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R.B. Altman. 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17: 520–525.

Warnes, G., G. Gorjanc, F. Leisch, and M. Man, 2011 genetics: Population Genetics. R package version 1.3.6. Avaliable at: http://CRAN.R-project.org/package=genetics / (verified 30 July 2012)

Weigel, K. a, G. de Los Campos, a I. Vazquez, G.J.M. Rosa, D. Gianola, and C.P. Van Tassell. 2010. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. J. Dairy Sci. 93: 5423–5435.

Whittaker, J. C., R. Thompson, and M. C. Denham, 2000 Marker-assisted selection using ridge regression. Genet. Res. 75: 249-252.

Windhausen, V.S., G.N. Atlin, J.M. Hickey, J. Crossa, J.-L. Jannink, M.E. Sorrells, B. Raman, J.E. Cairns, A. Tarekegne, K. Semagn, Y. Beyene, P. Grudloyma, F. Technow, C. Riedelsheimer, and A.E. Melchinger. 2012. Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 (Bethesda). 2: 1427–1436.

Wong, C. K., and R. Bernardo, 2008 Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theor. Appl. Genet. 116: 815-824.

Zhang Z., and T. Druet, 2010 Marker imputation with low-density marker panels in Dutch Holstein cattle. J. of Dairy Sci. 93: 5487–5494

Zhao, Y., M. Gowda, W. Liu, T. Würschum, H. Maurer, F. Longin, N. Ranc, and J. Reif. Accuracy of genomic selection in European maize elite breeding populations. TAG Theor. Appl. Genet. 124: 769–776.

Supplemental information

Methods

Optimal k value estimation for kNNI and SVDI: Optimal k values for kNNI and SVDI were estimated for the first replicate of each of the 15 datasets and these estimates were used for all remaining replicates. Optimal k values were estimated using 10-fold cross validation. For this procedure a set of k values: 1, 5, 10, 15, 20, 25 were chosen for initial evaluation. For each proposed value of k, a 10-fold cross validation was used to compute the accuracy in terms of median R_m^2. If the largest k value in the initial set of k values was found to be optimal, a new set of larger k values was evaluated. The values in the interval between two k values leading to the highest cross validation accuracy were then selected for the second round of k value evaluation. In this second round, the k value leading to the highest accuracy was considered to be optimal. This process was repeated until a k value leading to maximum cross validation accuracy was determined. To compute the cross validated accuracy, 1) 10 independent sets of non-missing data-points were identified, 2) set one data-points were masked, 3) either kNNI or SVDI was completed using the k value to be evaluated, and 4) steps 2 and 3 were repeated for all 10 sets. The median R_m^2 between the initial dataset and the dataset post-imputation of all 10 sets was used as the evaluation of cross validation accuracy.

Equivalent percent correct calculation: For each marker, 1001 marker genotype vectors, with the marker's MAF were simulated. Each vector had a

length of 1000. One of the vectors was selected as the true genotype, and the remaining 1000 were simulated to have different percent correct values, ranging from 0.01 to 100, with an interval of 0.1 between consecutive percent correct values. For each of the 1000 vectors with known percent correct values, R_m^2 was calculated. Then the vector with the R_m^2 value closest to the R_m^2 value for the marker of interest was identified, and that vector's known percent correct value was used as the equivalent percent correct value for the marker of interest.

Genomic selection accuracy calculations (continued): For the WW, SW, DTM and SRRW datasets the breeding values used for GS model training and validation consisted of best linear unbiased predictors (BLUPs) of the phenotypic values for the genotyped individuals. For the NAB dataset, an individual's phenotypic value *per se* was used as its breeding value. The GS accuracies for all marker set-imputation method combinations were calculated for a single trait with the SW, DTM, NAB, and SRRW datasets and for four traits with the WW dataset. (The traits that were used are listed in the section describing the original datasets). For more details on the BLUP calculations refer to Heffner et al. (2011) for the WW data and to Crossa et al. (2010) for the SW and DTM data. To compute BLUPs of the phenotypes for the SRRW data the mixed model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

was fit to the data. Y was the vector of phenotypic observations, $\boldsymbol{\beta}$ was the vector of site effects treated as fixed effects, $\mathbf{u}$ was the vector of genotype effects treated as random effects, $\mathbf{X}$ and $\mathbf{Z}$ were the design matrices relating the observations in

Y to **β** and **u**.

A 10-fold cross validation was used to compute GS accuracy. This consisted of 1) splitting the dataset into 10 sets, 2) training the model with 9 sets and predicting the remaining set, and 3) repeating steps one and two until predicted values have been calculated for all the individuals. The accuracy was defined as the Pearson's correlation between the breeding values estimated with phenotype and the genomic estimated breeding values (GEBVs). For all versions; NA0, NA20, NA50, and NA70, of a given dataset, individuals were assigned to specific sets that were held constant across all replicates, missing data levels, and traits in order to remove variation in predicted values that would arise due to sampling. This enabled direct comparison of the impact of the different imputation methods and missing data levels on the GS accuracy.

Ridge regression (RR) (Whittaker et al., 2000) and Bayesian Lasso (BL, de los Campos et al., 2009) were the two prediction models used for computing GS accuracies. For both RR and BL, marker effects were first estimated using the training set. These marker effect estimates and the genotypes of the validation individuals were used to calculate the GEBVs which were defined as the sum of each individual's marker effects. RR assumes that all marker effects are sampled from the same normal distribution with zero mean and variance that is estimated by maximum likelihood. With BL, the variance of the marker effect sampling distribution is unique for each marker. This leads to more and less shrinkage on small- and large-effect markers, respectively. We implemented RR in R (R

Development Core Team, 2011). The package *emma* (Kang et al., 2008) was used to estimate the variance components by maximum likelihood. BL was implemented in the R package *BLR* (de los Campos and Perez Rodriguez, 2010). The parameter values were set to those suggested by Perez et al. 2010 (Pérez et al., 2010). Marker effect estimations were based on 40,000 iterations of sampling after a burn in period of 20,000 iterations. Trace plots of the variance parameters were inspected to check for convergence.

Results

Effect of excluding sparse marker data on the genomic selection accuracy : In order to determine if markers with a large proportion of missing data should be included rather than filtered from the dataset, we assessed the effect of excluding sparse markers, those with a large proportion of missing data points, on the GS accuracy. Subsets of the NA70 versions of each dataset were used to calculate GS accuracies after imputation with each method. One of the two subsets contained markers that had up to 20% missing data per-marker before imputation, referred to as NA70-sub20. The second subset contained markers that had up to 50% missing data before imputation, referred to as NA70-sub50. The marker set containing markers with up to 70% missing data (which includes all the markers) is referred to as NA70-sub70. For comparison, the original datasets with no simulated missing data were also subsetted so they would contain the same markers as the NA70-sub20 and NA70-sub50, and NA70-sub70 datasets, these marker sets are referred to as NA0-sub20 and NA0-sub50,

NA0-sub70. The numbers of markers in each of the marker sets are listed in Table S1, and an example of the marker sets is illustrated in Figure S2. GS accuracy for each marker set-imputation method combination was calculated using RR. For each imputation method, the differences in GS accuracy between versions NA70-sub20, NA70-sub50, and NA70-sub70 were examined to determine how the GS accuracy is affected by including markers with over 20% missing values and with over 50% missing values after applying each imputation method. GS accuracy was also obtained using versions NA0-sub20, NA0-sub50, and NA0-sub70 to determine how the marker subsets affect the GS accuracy when the true genotypic data is known.

Fst calculation: Using each original dataset, individuals were classified into clusters using model based hierarchical agglomerative clustering described by Fraely and Raftery (2002) implemented using the R package *mclust* (Fraley et al. 2012). The multivariate normal mixture models evaluated were spherical equal volume, spherical unequal volume, diagonal equal volume and shape, diagonal equal volume, varying shape, diagonal varying volume, equal shape, and diagonal varying volume and shape. The number of clusters evaluated for each model was 1-15. For each dataset the optimal model and number of clusters was chosen according to the Bayesian information criterion. For all datasets the optimal model was diagonal varying volume, equal shape and the optimal number of clusters was 5, 3, 4, and 5 for the WW, SW, DTM, and NAB datasets respectively.

After individuals were classified into clusters, an Fst value for each marker was calculated to determine its level of differentiation due to genetic structure, or in other words, the amount its variance explained by population structure. Fst was calculated as:

$$\text{Fst} = \frac{\bar{p}^2 - \bar{P}^2}{\bar{P}(1 - \bar{P})}$$

where $\bar{p}^2$ is the weighted average of the squared allele frequency across subpopulations for one (arbitrary) allele, and $\bar{P}$ is the weighted average allele frequency across the subpopulations for that same allele (Weir and Cockerham, 1984).

Results

Optimal k values for k nearest neighbors and singular value decomposition imputation: The optimal k values for KNNI and SVDI for each dataset and dataset version are listed in Table S1. The optimal k values for KNNI were low, usually between 2 and 4 for most datasets and dataset versions. Compared to the other datasets, optimal k values for the DTM dataset were more than 10 times larger. The optimal KNNI k value for the SRRW dataset version NA70 was also disproportionally higher than the other dataset and dataset versions. The optimal k vales for SVDI varied depending on the dataset and always decreased as the level of missing data increased. It appeared that higher levels of LD between marker pairs and greater numbers of markers favored larger optimal k-values for SVDI.

Effect of excluding sparse marker data on the genomic selection accuracy: For all imputation methods, there was generally a sharp increase in GS accuracy across the NA70-sub20, and NA70-sub50 versions, and a slight increase across the NA70-sub50 and NA70-sub70 versions (Figure S3, panels B-F) indicating that excluding sparse marker data almost always lead to decreased accuracy especially when a more stringent percent missing threshold was used to filter markers. We found that filtering out the same marker sets when the true data was known had an even larger effect on the GS accuracy (Figure S3 panel A), indicating that marker density was a factor limiting the GS accuracy in these populations. Had marker density not been limiting, including sparse markers may not have lead to increased accuracy.

References

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, J.M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182: 375-85.

de los Campos, G., and P. Perez Rodriguez, 2010 BLR: Bayesian linear regression. R package versión 1.2. Avaliable at: http://CRAN.R-project.org/package=BLR/ (verified 30 July 2012).

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueño, J. L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, Y. Jianbing. V. Arief, M. Banziger, H.-J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713-724.

Fraley, C., and A.E. Raftery, 2002 Model-based clustering, discriminant analysis, and density estimation. J. Amer. Statist. Assoc. 97: 611-631.

Fraley, C., A.E. Raftery, T.B. Murphey, and L. Scrucca, 2012 mclust Version 4 for R: normal mixture modeling for model-based clustering, classification, and

density estimation technical report No. 597, Department of Statistics, University of Washington.

Heffner, E. L., J.-L. Jannink, and M. E. Sorrells, 2011 Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Gen. 4: 65-75.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M.J. Daly, E. Eskin. 2008 Efficient control of population structure in model organism association mapping. Genetics 178: 1709-1723.

Pérez, P., G. D. L. Campos, J. Crossa, D. Gianola, and G. de los Campos, 2010 Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. Plant Gen. 3: 106-116.

R Development Core Team, 2011 R: A Language and Environment for Statistical Computing, Vienna. Available at: http://www.r-project.org/ (verified 30 July 2012).

Weir, B. S., and C. C. Cockerham, 1984 Estimating F-statistics for the analysis of population structure. Evolution 6: 1358-1370.

Whittaker, J. C., R. Thompson, and M. C. Denham, 2000 Marker-assisted selection using ridge regression. Genet. Res. 75: 249-252.

Version NA0

	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m18	m19	m20
g1	1	0	1	-1	-1	1	-1	1	1	1	0	-1	-1	0	1	1	-1	-1	-1	0
g2	-1	0	-1	0	1	0	1	-1	0	0	-1	0	0	1	0	0	1	-1	1	-1
g3	1	-1	1	0	0	0	-1	1	1	1	1	1	0	1	0	1	1	1	-1	-1
g4	-1	1	0	-1	0	1	0	0	-1	1	1	-1	0	-1	0	0	1	1	0	-1
g5	1	1	-1	0	-1	-1	0	1	-1	-1	-1	-1	0	0	1	-1	-1	0	-1	0
g6	1	0	0	-1	-1	0	1	-1	0	0	-1	1	0	-1	1	1	-1	1	-1	-1
g7	0	0	-1	-1	1	1	-1	1	-1	0	-1	1	1	1	0	1	-1	1	-1	0
g8	1	0	0	-1	-1	1	-1	-1	-1	-1	0	-1	-1	-1	-1	1	-1	-1	0	-1
g9	0	-1	-1	1	-1	-1	0	1	0	1	1	-1	1	-1	0	0	0	0	0	0
g10	1	1	0	0	1	-1	1	0	0	0	1	0	1	1	0	0	0	-1	1	1
g11	0	0	1	1	-1	1	-1	0	0	1	1	1	-1	1	1	0	0	-1	0	0
g12	-1	1	1	-1	0	0	0	-1	-1	1	1	1	0	0	1	-1	0	0	-1	1
g13	1	-1	0	-1	0	0	1	-1	1	0	0	-1	0	1	0	-1	-1	0	1	0
g14	-1	-1	-1	1	0	1	-1	0	1	0	1	-1	-1	1	1	-1	0	1	-1	-1
g15	0	-1	0	-1	1	0	-1	1	0	0	1	-1	1	0	0	1	1	0	-1	0

Version NA20

	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m18	m19	m20
g1	1	0	1	-1	-1	1	-1	1	1	1	0	-1	-1	0	1	1	-1	-1	-1	0
g2	-1	NA	-1	0	1	0	1	-1	0	0	-1	0	0	NA	-1	0	1	NA	-1	-1
g3	1	-1	NA	-1	NA	-1	-1	1	1	1	1	1	0	1	0	NA	-1	1	-1	-1
g4	NA	1	0	-1	0	NA	-1	0	-1	NA	-1	-1	0	-1	0	0	1	1	0	-1
g5	1	1	-1	0	-1	-1	0	NA	-1	-1	-1	-1	0	0	1	-1	NA	-1	-1	0
g6	1	0	0	-1	NA	-1	NA	-1	0	0	-1	1	NA	-1	1	1	-1	1	-1	-1
g7	0	0	-1	-1	1	1	-1	1	-1	0	-1	NA	-1	1	0	1	-1	1	-1	0
g8	1	0	NA	-1	-1	1	-1	-1	-1	NA	-1	-1	-1	-1	NA	NA	-1	-1	0	NA
g9	NA	-1	-1	1	-1	NA	-1	1	0	1	1	-1	1	-1	0	0	0	NA	-1	0
g10	1	1	0	0	1	-1	NA	1	0	0	1	0	1	1	0	0	0	-1	1	1
g11	0	0	1	1	-1	NA	-1	0	0	1	1	1	-1	1	1	0	NA	-1	0	0
g12	-1	1	1	-1	0	0	0	-1	-1	1	1	1	0	0	1	-1	0	0	-1	1
g13	1	-1	0	-1	0	0	1	-1	1	0	0	-1	0	1	0	-1	-1	0	1	0
g14	-1	-1	NA	-1	0	1	-1	0	1	0	1	-1	NA	-1	1	-1	0	1	-1	-1
g15	0	-1	0	-1	1	0	-1	1	0	0	NA	-1	1	0	0	NA	-1	0	-1	0

Version NA50

	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m18	m19	m20
g1	NA	0	NA	NA	-1	1	-1	NA	1	1	NA	NA	-1	0	NA	1	NA	-1	NA	NA
g2	NA	0	-1	0	NA	0	NA	-1	NA	0	-1	0	0	NA	0	NA	1	-1	NA	NA
g3	1	-1	1	0	0	NA	-1	1	1	NA	1	NA	NA	1	NA	1	NA	NA	-1	NA
g4	-1	1	NA	NA	0	1	0	0	-1	1	1	-1	0	-1	0	0	1	1	0	-1
g5	NA	NA	-1	0	NA	-1	NA	1	NA	-1	-1	-1	NA	NA	1	NA	-1	0	NA	0
g6	1	0	0	-1	-1	0	1	NA	NA	NA	-1	1	0	-1	NA	1	NA	NA	NA	-1
g7	NA	NA	NA	NA	1	1	-1	1	NA	0	-1	1	1	NA	0	1	NA	1	-1	NA
g8	1	0	0	-1	-1	1	-1	-1	NA	-1	0	-1	NA	-1	-1	1	-1	-1	NA	-1
g9	NA	-1	-1	1	NA	NA	NA	1	0	1	1	-1	NA	-1	NA	NA	0	0	0	0
g10	1	NA	NA	NA	1	-1	1	0	NA	0	NA	NA	1	1	0	0	0	NA	1	NA
g11	0	0	1	1	-1	1	-1	0	0	NA	1	1	NA	NA	1	0	NA	-1	0	0
g12	NA	1	1	NA	0	0	NA	NA	-1	1	1	1	0	0	NA	-1	0	0	-1	NA
g13	NA	NA	0	-1	NA	NA	NA	-1	1	NA	NA	-1	NA	1	NA	NA	-1	NA	NA	0
g14	-1	-1	-1	NA	0	1	-1	0	1	0	1	-1	NA	1	1	-1	0	NA	-1	NA
g15	0	-1	0	NA	1	0	-1	1	NA	NA	1	NA	1	0	0	1	NA	0	-1	0

Version NA70

	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m18	m19	m20
g1	NA	NA	NA	-1	NA	NA	-1	1	1	NA	NA	-1	NA	0	NA	NA	-1	-1	NA	0
g2	NA	0	NA	NA	1	NA	1	NA	NA	NA	NA	NA	NA	1	NA	NA	1	-1	1	NA
g3	1	NA	1	NA	NA	0	-1	NA	1	NA	NA	NA	NA	1	0	NA	1	NA	-1	NA
g4	NA	NA	NA	-1	0	NA	NA	0	-1	1	NA	-1	0	-1	NA	0	NA	NA	NA	NA
g5	1	1	NA	NA	-1	-1	0	1	NA	NA	-1	-1	NA	0	-1	-1	-1	NA	NA	NA
g6	1	NA	0	NA	-1	NA	1	NA	0	NA	NA	NA	1	0	-1	NA	-1	1	NA	-1
g7	NA	NA	NA	-1	1	1	-1	NA	NA	0	NA	NA	1	1	NA	1	-1	NA	NA	0
g8	1	0	0	-1	NA	NA	NA	NA	-1	NA	0	-1	NA	NA	-1	1	NA	-1	NA	NA
g9	NA	NA	-1	NA	-1	NA	0	1	0	1	NA	NA	1	-1	NA	NA	0	NA	NA	NA
g10	1	1	NA	NA	NA	-1	NA	NA	NA	0	1	NA	NA	1	NA	NA	NA	-1	NA	NA
g11	0	NA	NA	1	-1	NA	-1	0	0	NA	NA	NA	NA	NA	NA	NA	0	-1	0	NA
g12	-1	NA	1	-1	NA	NA	0	NA	-1	1	NA	NA	0	0	-1	NA	0	NA	-1	1
g13	NA	-1	0	NA	0	0	1	NA	NA	NA	0	-1	NA	1	NA	NA	-1	0	NA	NA
g14	NA	NA	NA	NA	0	NA	NA	NA	1	NA	NA	NA	NA	1	1	-1	0	1	-1	NA
g15	0	NA	NA	-1	1	NA	-1	1	NA	NA	1	NA	1	0	0	NA	NA	0	-1	0

Figure S2.1: Illustration of example dataset versions NA20, NA50, and NA70

Simulated missing values are depicted in black. Rows (g1-g15) are individual genotypes and columns (m1-m20) are markers. Versions NA20, NA50, and NA70, have up to 20%, 50% and 70% missing data per marker respectively.

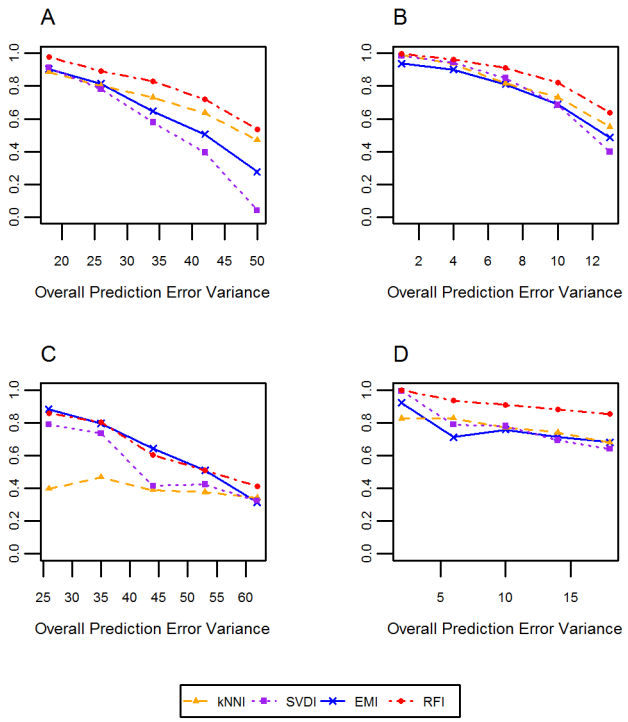


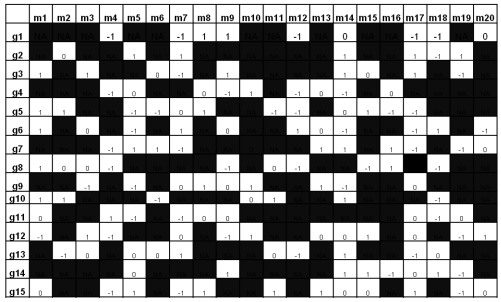
Figure S2.2: Relationship between the overall expected prediction error variance (PEV) and $\bar{\bar{R}}_i^2$

The median $\bar{\bar{R}}_i^2$ obtained for a given PEV value is plotted for each dataset: (A) Cornell winter wheat (WW), (B) CIMMYT elite spring wheat (SW), (C) CIMMYT drought tolerant maize (DTM), (D) North American barley (NAB). Each color and symbol represents a different imputation method: k-nearest neighbors imputation (kNNI, orange triangles), singular value decomposition imputation (SVDI, purple squares), random forest regression imputation (RFI, red circles), and expectation maximization imputation (EMI, blue crosses).

Figure S2.3: Illustration of the construction of marker sets used to determine the effect of excluding sparse marker data on the genomic selection accuracy.

Simulated missing values are depicted in black. Rows (g1-g15) are individual genotypes and columns (m1-m20) are markers. For each population the marker set version NA70 with up to 70% simulated missing data per marker was used and for each marker the percent missing was calculated. This marker set was then imputed with mean imputation, k-nearest neighbors imputation, singular value decomposition imputation, random forest imputation, and expectation maximization imputation. Markers in the imputed sets were then filtered based on their percent missing in version NA70 to create a subset of markers which had up to 20% missing: NA70-sub20, and up to 50% missing: NA70-sub50. For comparison, markers in the original marker set, NA0 were filtered based on their percent missing in version NA70.

Version NA70



	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m18	m19	m20
g1				-1			-1	1	1			-1		0			-1	-1		0
g2		0			1		1							1			1	-1	1	
g3	1		1			0	-1		1					1	0		1		-1	
g4				-1	0			0	-1	1		-1	0	-1		0				
g5	1	1			-1	-1	0	1			-1	-1		0	1	-1	-1			
g6	1		0		-1		1		0			1	0	-1			-1	1		-1
g7				-1	1	1	-1						1	1		1	-1		-1	0
g8	1	0	0	-1				1		0	-1				-1	1		1		
g9			1		-1		0	1	0	1			1	-1			0			
g10	1	1				-1				0	1		1		1			-1		
g11	0			1	-1		-1	0	0								0	-1	0	
g12	-1		1	-1			0		-1	1			0	0	1		0		-1	1
g13		-1	0		0	0	1				0	-1		1			-1	0		
g14				0					1					1	1	-1	0	1	-1	
g15	0			-1	1		-1	1			1		1	0	0		1		-1	0

Version NA70 after imputation

	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m18	m19	m20
g1	0.93	-0.02	0.96	-1	-0.99	0.95	-1	1	1	0.97	-0.03	-1	-0.99	0	0.97	0.97	-1	-1	-1.03	0
g2	-0.99	0	-1	-0.22	1	-0.01	1	-0.99	-0.07	-0.06	-0.99	-0.01	-0.09	1	-0.02	-0.01	1	-1	1	-0.99
g3	1	0.99	1	-0.01	-0.02	0	-1	0.97	1	0.96	0.38	0.96	-0.17	1	0	0.97	1	0.97	-1	-0.99
g4	-0.99	0.97	-0.05	-1	0	0.97	-0.07	0	-1	1	0.97	-1	0	-1	-0.07	0	-0.15	0.95	-0.01	-1
g5	1	1	-0.99	-0.01	-1	-1	0	1	-0.99	-1	-1	-1	-0.01	0	1	-1	-1	-0.01	-0.99	-0.01
g6	1	-0.01	0	-1	-1	-0.01	1	-0.99	0	-0.02	-0.99	1	0	-1	0.92	0.97	-1	1	-1	-1
g7	-0.01	-0.01	-1	-1	1	1	-1	0.96	-1	-0.02	-1	0.89	1	1	-0.01	1	-1	0.94	-1	0
g8	1	0	0	-1	-0.99	0.96	-0.99	-0.99	-1	-0.99	0	-1	-0.99	-0.99	-1	1	-1	-1	-0.01	-0.99
g9	-0.01	-0.99	-1	0.97	-1	-1	0	1	0	1	0.97	-0.99	1	-1	-0.17	-0.09	0	-0.01	-0.01	-0.01
g10	1	1	-0.12	-0.01	0.95	-1	0.97	-0.02	-0.01	0	1	-0.01	1	1	-0.01	-0.01	-0.01	-1	0.97	0.97
g11	0	-0.01	0.97	1	-1	0.97	-1	0	0	0.9	0.96	0.97	-0.99	0.97	0.96	-0.01	0	-1	0	-0.01
g12	-1	0.97	1	-1	-0.01	-0.01	0	-0.99	-1	1	0.96	0.97	0	0	1	-0.99	0	-0.01	-1	1
g13	0.95	-1	0	-1	0	0	1	-0.99	0.97	-0.03	0	-1	-0.02	1	-0.01	-0.99	-1	0	0.97	-0.01
g14	-1	-0.99	-0.95	0.96	0	0.97	-0.99	-0.02	1	-0.01	0.97	-0.99	-0.99	1	1	-1	0	1	-1	-1
g15	0	-0.99	-0.97	-1	1	-0.04	-1	1	-0.02	-0.01	1	-0.99	1	0	0	0.92	1	-0.01	-1	0

Version NA0

	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m18	m19	m20
g1	1	0	1	-1	-1	1	-1	1	1	1	0	-1	-1	0	1	1	-1	-1	-1	0
g2	-1	0	-1	0	1	0	1	-1	0	0	-1	0	0	1	0	0	1	-1	1	-1
g3	1	-1	1	0	0	0	-1	1	1	1	1	1	0	1	0	1	1	1	-1	-1
g4	-1	1	0	-1	0	1	0	0	-1	1	1	-1	0	-1	0	0	1	1	0	-1
g5	1	1	-1	0	-1	-1	0	1	1	-1	-1	-1	1	0	0	1	-1	-1	0	0
g6	1	0	0	-1	-1	0	1	-1	0	0	-1	1	0	-1	1	1	-1	1	-1	-1
g7	0	0	-1	-1	1	-1	1	-1	0	-1	1	1	0	1	-1	0	1	1	-1	0
g8	1	0	0	-1	-1	1	-1	-1	-1	-1	0	-1	-1	-1	-1	1	-1	1	0	-1
g9	0	-1	-1	1	-1	-1	0	1	0	1	1	-1	1	-1	0	0	0	0	0	0
g10	1	1	0	0	1	-1	1	0	0	0	1	0	1	1	0	0	0	-1	1	1
g11	0	0	1	1	-1	1	-1	0	0	1	1	1	-1	1	1	0	0	-1	0	0
g12	-1	1	1	-1	0	0	0	-1	-1	1	1	1	0	0	1	-1	0	0	-1	1
g13	1	0	0	1	-1	0	1	-1	0	0	-1	1	1	1	0	1	-1	0	1	-1
g14	-1	-1	-1	1	0	1	-1	0	1	0	1	-1	-1	1	1	-1	0	1	-1	-1
g15	0	-1	0	-1	1	0	-1	1	0	0	1	-1	1	0	0	1	1	0	-1	0

Filter markers based on the percent missing in Version NA70

NA70-sub50: Up to 50% missing

	m1	m5	m6	m7	m9	m14	m17	m18
g1	0.93	-0.99	0.95	-1	1	0	-1	-1
g2	-0.99	1	-0.01	1	-0.07	1	1	-1
g3	1	-0.02	0	-1	1	1	1	0.97
g4	-0.99	0	0.97	-0.07	-1	-1	-0.15	0.95
g5	1	-1	-1	0	-0.99	0	-1	-0.01
g6	1	-1	-0.01	1	0	-1	-1	1
g7	-0.01	1	1	-1	-1	1	-1	0.94
g8	1	-0.99	0.96	-0.99	-1	-0.99	-1	-1
g9	-0.01	-1	-1	0	0	-1	0	-0.01
g10	1	0.95	-1	0.97	-0.01	1	-0.01	-1
g11	0	-1	0.97	-1	0	0.97	0	-1
g12	-1	-0.01	-0.01	0	-1	0	0	-0.01
g13	0.95	0	0	1	0.97	1	-1	0
g14	-1	0	0.97	-0.99	1	1	0	1
g15	0	1	-0.04	-1	-0.02	0	1	-0.01

NA70-sub20: Up to 20% missing

	m14	m17
g1	0	-1
g2	1	1
g3	1	1
g4	-1	-0.15
g5	0	-1
g6	-1	-1
g7	1	-1
g8	-0.99	-1
g9	-1	0
g10	1	-0.01
g11	0.97	0
g12	0	0
g13	1	-1
g14	1	0
g15	0	1

NA0-sub50: Up to 50% missing

	m1	m5	m6	m7	m9	m14	m17	m18
g1	1	-1	1	-1	1	0	-1	-1
g2	-1	1	0	1	0	1	1	-1
g3	1	0	0	-1	1	1	1	1
g4	-1	0	1	0	-1	-1	1	1
g5	1	-1	-1	0	-1	0	-1	0
g6	1	-1	0	1	0	-1	-1	1
g7	0	1	1	-1	-1	1	-1	1
g8	1	-1	1	-1	-1	-1	-1	-1
g9	0	-1	-1	0	0	-1	0	0
g10	1	1	-1	1	0	1	0	-1
g11	0	-1	1	-1	0	1	0	-1
g12	-1	0	0	0	-1	0	0	0
g13	1	0	0	1	1	1	-1	0
g14	-1	0	1	-1	1	1	0	1
g15	0	1	0	-1	0	0	1	0

NA0-sub20: Up to 20% missing

	m14	m17
g1	0	-1
g2	1	1
g3	1	1
g4	-1	1
g5	0	-1
g6	-1	-1
g7	1	-1
g8	-1	-1
g9	-1	0
g10	1	0
g11	1	0
g12	0	0
g13	1	-1
g14	1	0
g15	0	1

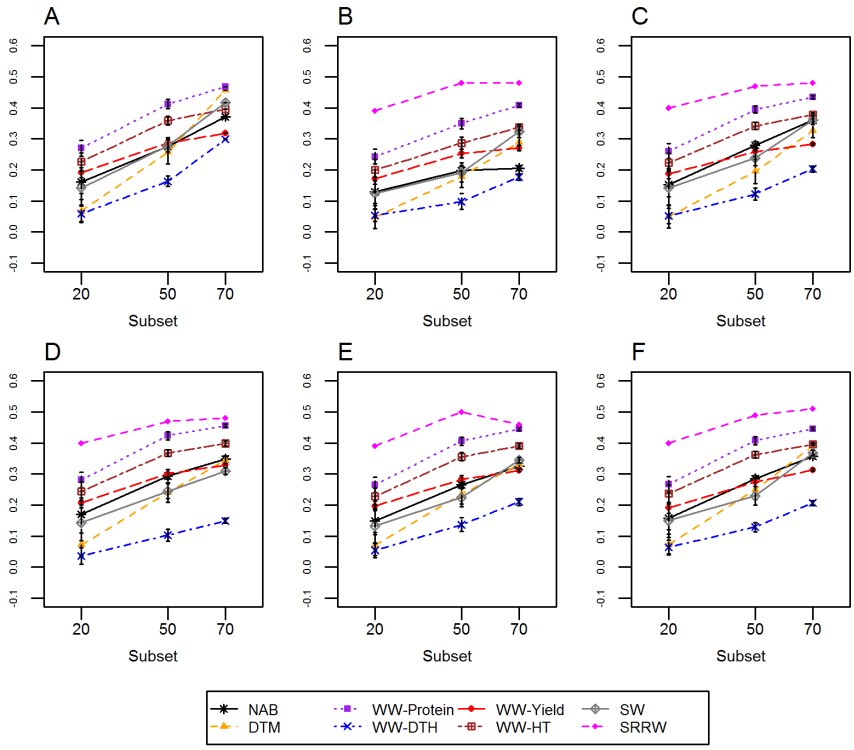


Figure S2.4: The effect of excluding sparse marker data on the genomic selection accuracy

Mean prediction accuracies obtained with different subsets of the NA70 dataset versions, which had up to 70% missing data per marker, are shown. The subsets compared were NA70-sub20, NA70-sub50, and NA70-sub70 (B-F), these marker sets contained markers with up to 20%, 50% and 70% missing values respectively. Prediction accuracies were also obtained with subsets of the NA0 data set version, which had up to 0% missing data per marker. These subsets were: NA0-sub20, NA0-sub50, and NA0-sub70 (A) and they consisted of the same set of markers as versions NA70-sub20, NA70-sub50, and NA70-sub70 respectively. The imputation methods used were (B) mean imputation (MNI), (C) k nearest neighbors imputation (kNNI), (D) singular value decomposition imputation (SVDI), (E) expectation maximization imputation (EMI), and (F) random forest imputation (RFI). In each panel prediction accuracies are shown for the population-traits: North American barley (NAB; black stars), CIMMYT drought tolerant maize (DTM; orange triangles), Cornell winter wheat (WW)-protein (purple squares), Cornell winter wheat-days to heading (WW-DTH; blue crosses), Cornell winter wheat (WW)-yield (red circles), Cornell winter wheat-height (WW-HT; brown squares), CIMMYT elite spring wheat (SW; grey open squares), stem rust resistant wheat (SRRW; pink diamonds). Error bars depict standard errors.

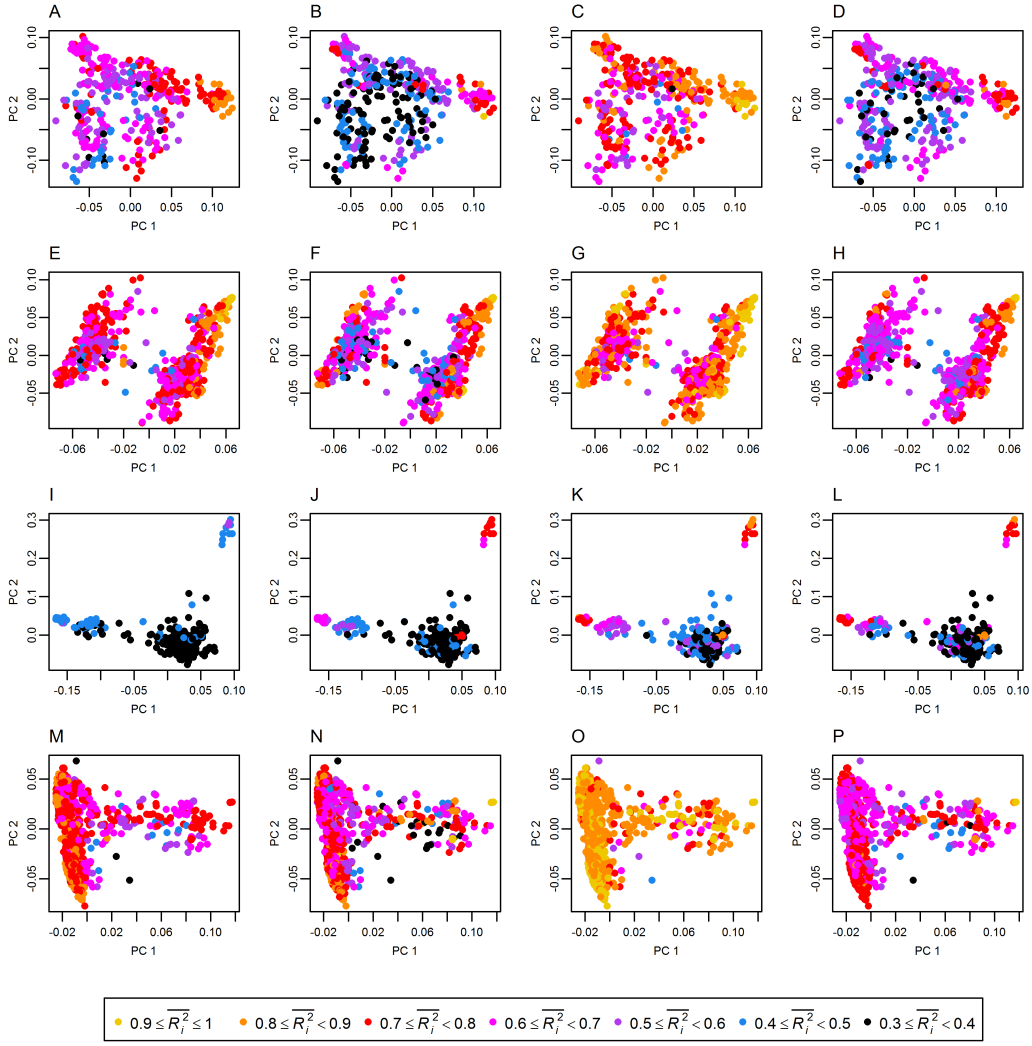


Figure S2.5: Heterogeneity of accuracies across population sub-groups

In each panel, principal component (PC) 2 *vs.* PC 1 of genotypic data is plotted to show population sub-groups. Each panel corresponds to a dataset-imputation method combination. Individuals are color coded according to their overall average imputation accuracy on an individual genotype basis, $\overline{\overline{R}_i^2}$. The colors yellow, orange, red, pink, purple, blue, and black correspond to $0.9 \leq \overline{\overline{R}_i^2} \leq 1$, $0.8 \leq \overline{\overline{R}_i^2} < 0.9$, $0.7 \leq \overline{\overline{R}_i^2} < 0.8$, $0.6 \leq \overline{\overline{R}_i^2} < 0.7$, $0.5 \leq \overline{\overline{R}_i^2} < 0.6$, $0.4 \leq \overline{\overline{R}_i^2} < 0.5$, $0.3 \leq \overline{\overline{R}_i^2} < 0.4$ respectively. Panels A-D correspond to the Cornell winter wheat (WW) data imputed with k nearest neighbors imputation (kNNI; A), singular value decomposition imputation (SVDI; B), random forest imputation (RFI, C), expectation maximization imputation (EMI, D). Panels E-H correspond to the CIMMYT elite spring wheat data imputed with kNNI (E), SVDI (F), RFI (G), EMI (H). Panels I-L correspond to the CIMMYT drought tolerant maize data imputed with kNNI (I), SVDI (J), RFI (K), EMI (L). Panels M-P correspond to the North American barley data imputed with kNNI (M), SVDI (N), RFI (O), EMI (P).

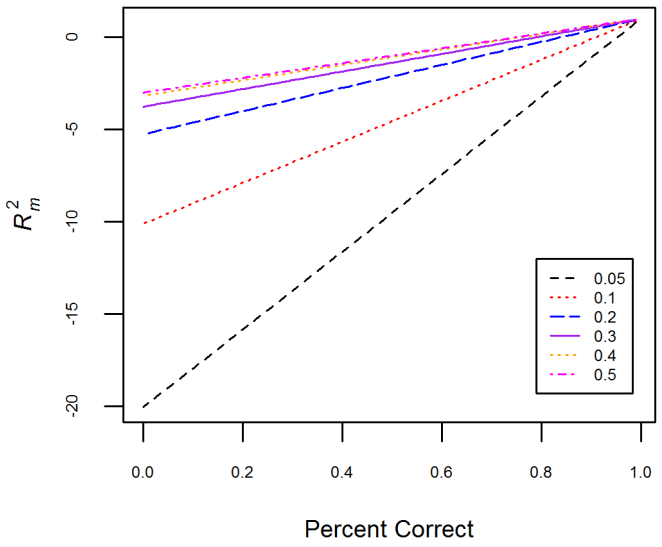


Figure S2.6: The relationship between imputation accuracy measured as R_m^2 and measured as percent correct for different minor allele frequencies

For each minor allele frequency value: 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5, which is depicted in black, red, blue purple ,orange and magenta respectively, the relationship between the R_m^2 and the percent correct is shown.

Table S2.1: Optimal k values for KNNI† and SVDI‡ used across all replicates

Dataset§	Version¶	KNNI	SVDI
WW	NA20	3	78
	NA50	3	51
	NA70	4	19
SW	NA20	2	88
	NA50	3	89
	NA70	4	33
DTM	NA20	51	10
	NA50	39	9
	NA70	49	5
NAB	NA20	4	190
	NA50	4	117
	NA70	4	50
SRRW	NA20	3	81
	NA50	4	36
	NA70	66	1

†KNNI: k-nearest neighbors imputation
‡SVDI: singular value decomposition imputation
§WW: Cornell winter wheat, SW: CIMMYT elite spring wheat, DTM: CIMMYT drought tolerant maize, NAB: North American barley, SRRW: CIMMYT stem rust resistant wheat
¶NA20: up to 20% missing data per marker, NA50: up to 50% missing data per marker, NA70: up to 70% missing data per marker

Table S2.2: Description of datasets used to test the effect of excluding sparse marker data on the genomic selection accuracy

Dataset†	Version‡	Mean number of markers
WW	NA70-sub20	20
	NA70-sub50	119
	NA70-sub70	1158
SW	NA70-sub20	23
	NA70-sub50	112
	NA70-sub70	1279
DTM	NA70-sub20	18
	NA70-sub50	99
	NA70-sub70	1135
NAB	NA70-sub20	37
	NA70-sub50	185
	NA70-sub70	2146
SRRW	NA70-sub20	34
	NA70-sub50	169
	NA70-sub70	2014

†WW: Cornell winter wheat, SW: CIMMYT elite spring wheat, DTM: CIMMYT drought tolerant maize, NAB: North American barley, SRRW: CIMMYT stem rust resistant wheat
‡Version NA70-sub20: up to 70% missing data per marker was simulated, and markers were discarded if they had over 20% missing data. Version NA70-sub50: up to 70% missing data per marker was simulated, and markers were discarded if they had over 50% missing data. Version NA70-sub70: up to 70% missing data per marker was simulated, and no markers were discarded.

CHAPTER 3

EVALUATION OF GENOMIC PREDICTION METHODS FOR FUSARIUM HEAD BLIGHT RESISTANCE IN WHEAT[3]

Abstract

Fusarium head blight (FHB) resistance is quantitative and difficult to evaluate. Genomic selection (GS) could accelerate FHB resistance breeding. We used U.S. cooperative FHB wheat nursery data to evaluate GS models for several FHB resistance traits including deoxynivalenol (DON) levels. For all traits we compared the models: ridge regression (RR), Bayesian Lasso (BL), reproducing kernel Hilbert spaces (RKHS) regression, random forest (RF) regression, and multiple linear regression (MLR) (fixed effects). For DON, we evaluated additional prediction methods including bivariate RR models, phenotypes for correlated traits, and RF regression models combining markers and correlated phenotypes as predictors. Additionally, for all traits, we compared different marker sets including genomewide markers, FHB quantitative trait loci (QTL) targeted markers, and both sets combined. Genomic selection accuracies were always higher than MLR accuracies, RF and RKHS regression were often the most accurate methods, and for DON, marker plus trait RF regression was more

---

[3] Originally published as Rutkoski, J., J. Benson, Y. Jia, G. Brown-Guedira, J.-L. Jannink, & M. Sorrells. 2012. Evaluation of genomic prediction methods for fusarium head blight resistance in wheat. Plant Gen., 5:51-61.

accurate than all other methods. For all traits except DON, using QTL targeted markers alone led to lower accuracies than using genomewide markers. This study indicates that cooperative FHB nursery data can be useful for GS, and prior information about correlated traits and QTL could be used to improve accuracies in some cases.

Abbreviations

BL, Bayesian Lasso; BLUP, best linear unbiased predictor; CV1, fivefold cross-validation; CV2, cross-validation across years; DArT, Diversity Array Technology; DON, deoxynivalenol; FDK, Fusarium damaged kernels; FHB, Fusarium head blight; GEBV, genomic estimated breeding value; GM, genomewide DArT markers only; GS, genomic selection; HD, days to heading; INC, incidence; ISK, incidence severity, and kernel quality index; LD, linkage disequilibrium; MAS, marker-assisted selection; MLR, multiple linear regression; NUWWSN, northern uniform winter wheat scab nursery; PEBV, phenotype-based estimate of breeding value; QTL, quantitative trait loci; RF, random forest; RKHS, reproducing kernel Hilbert spaces; RR, ridge regression; SEV, severity; SSR, simple sequence repeat; TM, QTL targeted SSR markers only; TM+GM QTL targeted SSR markers and genomewide DArT markers; USFHBN, uniform southern Fusarium head blight nursery

Introduction

Fusarium head blight (FHB) has been the most devastating plant disease affecting U.S. agriculture during the last decade (Windels, 2000). It is estimated that in the United States between 1998 and 2000 economic losses due to FHB

reached US$2.7 billion (Nganje et al., 2004). Fusarium head blight is primarily caused by the fungal pathogen, *Fusarium graminearum* Schwabe, which attacks the spikes of wheat (*Triticum aestivum* L.) and barley (*Hordeum vulgare* L.). The pathogen causes kernels to be shriveled and discolored and also produces the mycotoxin deoxynivalenol (DON), which can be toxic to humans and animals (Pestka and Smolinski, 2005) and can render grain unmarketable for human or animal consumption depending on the DON level.

Aside from improved phenotyping strategies including cooperative phenotyping and selection based on correlated traits, marker-assisted breeding methods are being pursued. To enable marker-assisted selection (MAS), over 40 quantitative trait loci (QTL) mapping studies have been conducted for FHB resistance and have identified over 200 QTL that are distributed across every chromosome (reviewed by Liu et al., 2009, and Buerstmayr et al., 2009). However, a handful of QTL have been validated across studies, and primarily one major QTL, *Fhb1*, that has been found to reduce disease by 20 to 25% on average (Pumphrey et al., 2007), has been the target for MAS aimed at improving FHB resistance levels (Anderson et al., 2007). The alleles known to confer FHB resistance at *Fhb1* and many other validated resistance loci with relatively large effects originate from Chinese sources and are at a very low frequency in North American germplasm (Sneller et al., 2010; Bernardo et al., 2011). To avoid linkage drag associated with introgressing resistance alleles from alien sources, the objective of many North American wheat breeders is to improve levels of

resistance based on existing variation in the native germplasm. Resistance from non-Chinese sources appears to be distinct and is conferred by many small effect loci originating from various different parents (Gosman et al., 2007; Buerstmayr et al., 2008; Sneller et al., 2010; Miedaner et al., 2010). Therefore, conventional MAS strategies to improve native FHB resistance in North American germplasm have not been used.

A new marker-assisted breeding method called genomic selection (GS) (Meuwissen et al., 2001) has great potential for use in crop plants (reviewed by Lorenz et al., 2011) for quantitative trait improvement and is expected to be more effective than MAS in many cases. Genomic selection is already routinely used in cattle (*Bos taurus*) breeding and should become an important tool in plant breeding because it can lead to greater gain from selection per unit of time and cost compared to phenotypic selection (Heffner et al., 2010). The accuracy of GS models for a range of quantitative traits has been demonstrated in various studies involving populations derived from biparental crosses (Lorenzana and Bernardo, 2009; Heffner et al., 2011a) and sets of breeding lines representative of a single breeding program (Crossa et al., 2010; Heffner et al., 2011b). However, few studies (Asoro et al., 2011; Lorenz et al., 2012) have evaluated GS across multiple breeding programs. One of these studies (Lorenz et al., 2012) found GS to be promising for FHB resistance in barley using cooperative nursery data. Whether this will be true for wheat requires validation.

Using cross-validation, accuracies of prediction methods can be compared

and the merits of incorporating prior knowledge about loci or correlated traits in the prediction models can be evaluated. Although GS requires minimal prior information about traits to be predicted, such information is available for traits such as FHB resistance and could be useful for improving prediction accuracy. Information about important loci affecting the trait could be used to help select markers for genotyping. In addition, component and/or correlated traits for the trait of interest could be included along with markers in one prediction model to reduce the cost per breeding cycle for traits such as DON whose evaluation may be more costly than both phenotyping a correlated trait and genotyping.

With this study we aim to (i) determine the potential utility of using GS as a tool to improve FHB resistance in wheat using cooperative nursery data involving wheat germplasm across the United States for GS modeling, (ii) compare the relative accuracy of several different marker-based prediction models to identify the most promising models, (iii) compare prediction accuracies achieved using genomewide markers, QTL targeted markers, or both types combined, and (iv) assess the utility of combining correlated trait measurements and markers in prediction models for resistance to DON accumulation and determine if these combined marker–trait prediction models are more accurate than using correlated traits alone.

Materials and Methods

Phenotypic data

The breeding lines used in this study consisted of 322 lines from 15 public

and three private breeding programs across the eastern United States and Canada that were evaluated in the 2008, 2009, and 2010 northern uniform winter wheat scab nursery (NUWWSN), the 2008 and 2009 preliminary NUWWSN, and the 2008 and 2009 uniform southern Fusarium head blight nursery (USFHBN). Of these 322 lines, 170 that had genotypic data and phenotypic data for all traits to be analyzed were used for prediction model evaluation. The NUWWSN, preliminary NUWWSN, and USFHBN were conducted under the coordination of the U.S. wheat and barley scab initiative whose aim is to develop control measures against FHB. Each nursery cooperator submits his or her breeding materials for evaluation and conducts an inoculated FHB trial at his or her location. The phenotypic data from the nurseries along with a list of the locations and cooperators involved is available at http://scabusa.org/publications.html#pubs_uniform-reports. According to a study by Benson and Brown-Guedira (2012) examining structure and diversity of these cooperative nurseries, subpopulation substructure in this germplasm is minimal and the effective population size (N_e) is estimated to be 45.

Each year–nursery consists of a set of locations, and lines within a year–nursery are evaluated across all locations within that year–nursery. Except for the checks 'Ernie', 'Truman', 'Freedom', and 'Pioneer 2545', on average lines were evaluated in two nursery–years and 11 year–nursery–locations. The phenotypic evaluations were conducted slightly differently depending on the location. The field design was a randomized complete block. The number of replications

ranged from two to six. Plot sizes were one, two, or four 1-m row(s). Artificial epidemics of FHB were created either by spreading diseased corn kernels throughout the plots before flowering or by spray inoculation of plots at 50% anthesis using a spore suspension (Gilbert and Woods, 2006).

The phenotypes evaluated in the nurseries were incidence (INC), severity (SEV), Fusarium damaged kernels (FDK), incidence, severity, and, kernel quality index (ISK) (Kolb and Boze, 2003), DON content of the grain, and days to heading (HD). Incidence is a visual measure of percentage of heads showing disease symptoms in a plot and is a measure of resistance to initial infection. Severity is visually measured on infected spikes as the percent of the spike showing symptoms and is a measure of resistance to fungal spread from a point of initial infection. Fusarium damaged kernels is measured on threshed grains and is a visual estimate of the percentage of kernels showing symptoms. Incidence, severity, and, kernel quality index is an index that combines SEV, INC, and FDK scores using the formula ISK = (0.3 INC(%)) + (0.3 SEV(%)) + (0.4 FDK(%)). And DON is the milligrams per kilogram measurement of toxin levels present in a 100-g sample of grain using an enzyme linked immunosorbent assay (Casale et al., 1988) or gas chromatography–electron capture (Pathre and Mirocha, 1977).

Genotypic data

For each entry, DNA was extracted from single seedlings using a cetyltrimethylammonium bromide extraction protocol described by Pallotta et al. (2003). The DNA was sent to Triticarte (http://www.triticarte.com.au) for whole-

genome genotyping using Diversity Array Technology (DArT) markers (Akbari et al., 2006). A total of 2402 polymorphisms were detected. Entries were also genotyped with simple sequence repeat (SSR) markers targeted to important FHB resistance QTL as described by Benson and Brown-Guedira (2012). A subset of eight SSR markers targeted to five QTL (Table 2.1) was selected to be included in the analysis.

Table 3.1: Markers included in the marker sets TM and TM+GM†

Marker	Chromosome	QTL‡ name	Reference
gwm157	2DL	*QFhs.nau-2DL* and/or *QFhs.crc-2D*	(Jiang et al., 2007a; Jiang et al., 2007b)
gwm539	2DL	*QFhs.nau-2DL* and/or *QFhs.crc-2D*	(Jiang et al., 2007a; Jiang et al., 2007b)
gwm533	3BS	*Fhb1*	(Zhou et al., 2002)
gwm493	3BS	*Fhb1*	(Anderson et al., 2007)
wmc152	6BS	*Fhb2*	(Cuthbert et al., 2007)
wmc238	4BS		(Somers et al., 2003)
barc117	5AS	*Qfhs.ifa-5A* and/or *Qfhs.umc-5A*	(Chen et al., 2006)
gwm304	5AS	*Qfhs.ifa-5A* and/or *Qfhs.umc-5A*	(Buerstmayr et al., 2002)

†TM, quantitative trait loci (QTL) targeted simple sequence repeat (SSR) markers only. TM+GM, QTL targeted SSR markers and genomewide Diversity Array Technology (DArT) markers combined.
‡QTL, quantitative trait loci.

Of these QTL, *Fhb1* has been a target of MAS in this germplasm. Each SSR allele was converted into a binary variable, resulting in 38 total variables. Three sets of markers were then constructed: (i) genomewide DArT markers only (GM), (ii) QTL targeted SSR markers only (TM), and (iii) QTL targeted SSR markers and genomewide DArT markers (TM+GM). Each of these three marker sets was used in each prediction model and across all traits. Linkage disequilibrium (LD) between the DArT markers and some of the SSR alleles was low, indicating that

the QTL targeted SSR markers captured variation at loci that was not captured by the DArT markers. In all marker sets, missing variables were imputed with the mean value for a particular variable.

Phenotype-based breeding value estimation

Using all the phenotypic data available for each nursery–year, a mixed effects model with sites as fixed effects and entries as random effects was fit using the *lme4* package (Bates et al., 2009) implemented in R (R Development Core Team, 2010) to calculate the best linear unbiased predictors (BLUPs) for each entry for each trait. The response variable was the average trait measurement for each entry within each site. For DON, a log transformation of the response variable was used to normalize the data before calculating BLUPs. Sites were defined as a unique year–location–nursery combination. The BLUPs were then used as the phenotype-based estimates of breeding values (PEBVs) for the subsequent model training and validation steps. Because BLUPs have reduced variance relative to the true breeding values (Garrick et al., 2009), we expect our results using BLUPs to train the prediction models to yield conservative results relative to what could be achieved from using de-regressed and weighted BLUPs. Proper de-regression and weighting of BLUPs from data collected using heterogeneous phenotyping methods, such as those used to collect the data used in this study, is an important area of investigation but is beyond the scope of this paper.

Using the same mixed model described above, BLUPs were also calculated

within years and within the following two-year combinations—2008 and 2009, 2008 and 2010, and 2009 and 2010—for all traits except DON because of insufficient 2010 data. These BLUP calculations were used as the PEBVs for the calculation of the across-year cross-validation prediction accuracies.

Prediction models

Genomic selection models and multiple linear regression (MLR) models were compared. The GS models tested were ridge regression (RR) BLUP (Meuwissen et al., 2001; Whittaker et al., 2000), Bayesian Lasso (BL) (de los Campos et al., 2009b; Park and Casella, 2008), reproducing kernel Hilbert spaces (RKHS) (Gianola et al., 2006) regression, and random forest (RF) (Breiman, 2001) regression. For a detailed description of RR, BL, and RKHS methods refer to Lorenz et al. (2011). For a description and comparative study of RR, BL, RKHS, and RF refer to Heslot et al. (2012). The GS models were tested with each of the three marker sets: GM, TM, and TM+GM. We also tested five bivariate RR models for DON using the TM+GM set where the model was trained with DON values in addition to either SEV, INC, FDK, or ISK values. For each marker-based prediction method genomic estimated breeding values (GEBVs) were calculated for each individual.

The MLR models for prediction were constructed slightly differently depending on the marker set used. The MLR models tested for each trait using the GM marker set and the GM+TM marker set involved two stages for each prediction: (i) association analysis and (ii) MLR using a subset of *k* markers with

the lowest p-values. In the TM marker set, all 38 SSR alleles converted to a binary format were used in MLR. With the GM marker set, before association analysis, markers with minor allele frequency less than 0.05 were removed and the LD tagSNP function, based on the algorithm described by Carlson et al. (2004) as implemented in JMP Genomics 5.0 (SAS Institute, 2010), was used to select nonredundant markers defined as those with r^2 values less than 0.75. This threshold level led to an adequate reduction of redundant markers especially for markers present on the alien translocations present on chromosomes 1B and 2B. The filtering also reduced the multicolinearity between the predictors to be fit in the MLR prediction model. The filtered GM marker set consisted of 900 markers. The filtered TM+GM marker set consisted of the 900 nonredundant DArT markers and all 38 SSR alleles. The association analysis stage was conducted using the package *emma* (Kang et al., 2008) implemented in R (R Development Core Team, 2010) to fit a mixed effects model testing each marker's association with the trait while correcting for multiple levels of relatedness (Yu et al., 2006). In each mixed effects model, a vector of the PEBVs was used as the response variable, a vector of the marker values at a particular locus was a fixed effect, a matrix (**Q**) of the first two principal components of the genotype matrix was a fixed effect used to correct for population structure, and a marker-based kinship matrix (**K**) was a random effect used to correct for family structure. To ensure that the **K** and **Q** matrix were adequately correcting for population and family structure, q-q plots of the p-values for the marker effects were examined to

ensure that the p-values were uniformly distributed. In stage two, *k* markers with the lowest p-values were selected as the explanatory variables in a fixed-effect model where the vector of the PEBVs was the response variable. The marker effects measured with the fixed-effects model were then used to calculate the predicted breeding values of individuals in the validation set. Multiple linear regression models with k = 5, 10, 15, 20, and 25 were tested to determine the optimal number of predictors to use in MLR. With the TM marker set, MLR was conducted as in stage two described above, but in this set, all 38 SSR alleles, targeted to five previously validated QTL regions, were used as the explanatory variables in a fixed effect model.

For RR and BL, marker effects were first estimated using the training set, and GEBVs of individuals in the validation set were calculated as the sum of each individual's marker effects. Ridge regression assumes that all marker effects are sampled from the same normal distribution with zero mean and variance that is estimated by maximum likelihood. For a more through description of RR refer to Whittaker et al. (2000) and Piepho (2009). With BL, the variance of the marker effect sampling distribution changes from marker to marker, forcing more and less shrinkage on small- and large-effect markers, respectively. For a more detailed description of BL refer to Pérez et al. (2010). We implemented RR in R (R Development Core Team, 2010) and used the package *emma* (Kang et al., 2008) to obtain maximum likelihood estimates of the variance components. We implemented BL using the R package *BLR* (de los Campos and Perez Rodriguez,

2010) using the parameter values suggested by Pérez et al. (2010). Marker effect estimations were based on 50,000 iterations of sampling after a burn-in period of 10,000 iterations. Trace plots of the variance parameters were inspected to ensure convergence was reached.

In RKHS regression, genetic values are assumed to be sampled from a normal distribution with zero mean and with a covariance structure proportional to a kernel matrix that is calculated by applying a kernel function to the marker data. For a more detailed description of RKHS refer to de los Campos et al. (2009a). We implemented RKHS regression in R (R Development Core Team, 2010) using functions adapted from those provided in the supplemental data of Crossa et al. (2010) We also used the parameter values suggested by Crossa et al. (2010). Genomic estimated breeding value estimations were based on 30,000 iterations of sampling after a burn-in period of 5000 iterations. Trace plots of the variance parameters were inspected to ensure convergence was reached.

Random forest regression (Breiman, 2001) uses an ensemble of multiple decision trees for prediction. Each tree is grown using a bootstrap sample of training individuals and markers are used as the node-splitting variables. Predicted values produced by each individual tree are averaged to obtain a single prediction. For a more detailed description of RF regression refer to Breiman (2001) and González-Recio and Forni (2011). We implemented RF regression using the R package *randomForest* (Liaw and Wiener, 2002). For each prediction we set the number of trees to 500. For DON we tested RF regression using ISK in

addition to markers as the predictors.

For the multivariate ridge regression model, a multitrait animal model $\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}$ was fit in which, for *n* individuals and m traits, **Y** is a vector *n* × *m* elements long and composed of n subvectors, each recording the observations for the m traits of each individual. **X** is the design matrix associating observations to the fixed effects **β**, **Z** is the design matrix allocating the observations to the individuals, **u** is an *n* × *m* long vector of breeding value random effects, and **ε** is an *n* × *m* long vector of residual errors with zero mean and *m* × *m* covariance matrix Σ_{ε}.

The variance of y is $\mathbf{ZZ}^{\mathrm{T}} \otimes \Sigma_{\mathrm{g}} + \mathbf{I} \otimes \Sigma_{\varepsilon}$, in which $\boldsymbol{\Sigma}_{\mathrm{g}}$ is an *m* × *m* covariance matrix of the genetic effects of lines. The elements of **u** were estimated using the software ASREML (Gilmour et al., 2009).

Cross-validation accuracy calculation

Accuracy (*r*) was defined as the Pearson's correlation between the PEBVs and the GEBVs calculated using cross-validation. For each trait–model–marker set combination two different cross-validation schemes were used. The first scheme (fivefold cross-validation [CV1]) consisted of a global fivefold cross-validation using the overall PEBVs for each entry as the phenotype-based estimate of the true breeding value. In this scheme the entries were randomized and then divided into five sets. Four of the five sets were used for model training, and this model was then used to calculate the GEBVs of the remaining set. In each

fold of the cross-validation, there were 136 individuals used for model training and 34 individuals used for validation. After GEBVs were calculated for all individuals the means and standard errors of the prediction accuracy were calculated using bootstrapping (Efron, 1979). Specifically, a bootstrap sample of the 170 total individuals was drawn and the correlation of the PEBVs and GEBVs was computed and saved. This was repeated 1000 times to obtain a distribution of accuracies. The mean of the distribution was used as the estimate of the mean accuracy and the standard deviation of the distribution was used as the standard error of the mean. For a review of this methodology refer to Efron and Tibshirani (1986). For DON, we also used this bootstrap method to calculate the means and standard errors of the Pearson's correlation between DON and INC, SEV, FDK, and ISK.

The second cross-validation scheme (cross-validation across years [CV2]) consisted of a single cross-validation across years where the PEBVs from entries evaluated in 2009 and 2008 were used to predict the PEBVs of the entries evaluated in 2010, PEBVs from entries evaluated in 2008 and 2010 were used to predict PEBVs from entries evaluated in 2009, and PEBVs from entries evaluated in 2009 and 2010 were used to predict PEBVs from entries evaluated 2008. No entries overlapped between training and validation sets. The number of entries used for training and validation respectively was 109 and 62 when 2008 was the validation set, 92 and 79 when 2009 was the validation set, and 141 and 30 when 2010 was the validation set. The average across year prediction accuracy per

trait–model combination was the mean across the three cross-validations, and the standard error was computed as the standard deviation/3½. Cross-validation across years was conducted for all traits except DON due to a lack of 2010 data.

Statistical testing for differences between models

To test for differences among CV1 means for each trait–model set combination we compared the 95% confidence intervals for each of the pairwise combinations of models tested for a given trait and marker set. Each confidence interval was computed using the bootstrap mean and standard deviation estimates described in the previous section. The 95% confidence interval was defined as the mean ± 1.96 × standard error. A pair of accuracies was considered significantly different if their confidence intervals did not overlap. For each trait we tested for differences among CV2 means for each of the prediction methods using an ANOVA. If we found the effect of a prediction model to be significant we conducted a Tukey's multiple comparisons of means test using a 95% familywise confidence interval to detect differences among pairs of mean prediction model accuracies.

Assessment of unintentional prediction of maturity

Because the development of FHB symptoms is sensitive to the proper timing of inoculation, using the TM+GM marker set we determined if the prediction models trained using PEBVs for SEV, INC, FDK, and ISK could predict the HD PEBVs of the validation set to ensure that the models were not capturing "passive" resistance due to maturity. Correlations between GEBVs for resistance

and PEBVs for HD were calculated to determine if they were nonzero, which would indicate that the prediction models for resistance may incorporate some level of passive resistance related to disease escape.

Results

Global and across-year cross-validated accuracies

Mean prediction accuracies ± standard errors for two different cross-validation schemes, CV1 and CV2, for each model, marker set, and trait (except for DON, where only CV1 was used) are reported in Table 3.2. For each trait–marker set combination, statistical comparisons between prediction model accuracies are also reported in Table 2.2. Differences between prediction models were clearer with the CV1 results compared to CV2 results because CV1 prediction accuracies had smaller standard errors. However, very few pairwise comparisons of prediction models for a given trait–marker set were significantly different. Significant differences in accuracy between different GS models were only detected for seven of the 35 different trait–marker set–cross-validation scheme combinations (Table 3.2).

Linear prediction model accuracies

Results from the MLR optimization step to determine how many predictors to use for each trait when using the GM or GM+TM marker sets are reported in Table 3.3.

Table 3.2: Means and standard errors of cross-validated prediction accuracies for all traits calculated using fivefold cross-validation (CV1) and cross-validation

across years (CV2). For each trait three different marker sets and five different prediction models are compared.

Trait†	Cross-validation scheme	Marker set‡	RF§	RKHS	RR	BL	MLR
HD	CV1	GM	0.387 ± 0.072 ab¶	0.403 ± 0.069 a	0.26 ± 0.083 ab	0.288 ± 0.083 ab	0.113 ± 0.07 b
		TM+GM	0.37 ± 0.074 ab	0.41 ± 0.075 a	0.271 ± 0.08 ab	0.247 ± 0.085 ab	0.107 ± 0.07 b
		TM	0.317 ± 0.081 a	0.33 ± 0.083 a	0.13 ± 0.079 a	0.104 ± 0.088 a	0.204 ± 0.068 a
	CV2	GM	0.411 ± 0.077 a	0.408 ± 0.103 a	0.432 ± 0.101 a	0.397 ± 0.002 a	0.008 ± 0.089 a
		TM+GM	0.414 ± 0.075 a	0.399 ± 0.094 a	0.436 ± 0.101 a	0.430 ± 0.099 a	0.008 ± 0.089 a
		TM	0.079 ± 0.046 a	0.248 ± 0.121 a	0.29 ± 0.01 a	0.283 ± 0.098 a	0.202 ± 0.052 a
FDK	CV1	GM	0.423 ± 0.065 a	0.426 ± 0.067 a	0.463 ± 0.063 a	0.412 ± 0.066 a	0.338 ± 0.067 a
		TM+GM	0.455 ± 0.06 a	0.398 ± 0.071 a	0.376 ± 0.064 a	0.399 ± 0.064 a	0.343 ± 0.068 a
		TM	0.19 ± 0.076 a	0.064 ± 0.078 a	0.033 ± 0.083 a	0.046 ± 0.088 a	0.015 ± 0.074 a
	CV2	GM	0.41 ± 0.077 a	0.389 ± 0.1 a	0.349 ± 0.167 a	0.307 ± 0.047 a	0.272 ± 0.136 a
		TM+GM	0.379 ± 0.092 a	0.398 ± 0.114 a	0.353 ± 0.164 a	0.286 ± 0.165 a	0.263 ± 0.079 a
		TM	0.079 ± 0.046 a	0.008 ± 0.03 a	0.006 ± 0.026 a	0.006 ± 0.019 a	−0.023 ± 0.052 a
ISK	CV1	GM	0.548 ± 0.05 a	0.542 ± 0.049 a	0.543 ± 0.055 a	0.455 ± 0.059 a	0.4 ± 0.064 a
		TM+GM	0.555 ± 0.051 a	0.56 ± 0.051 a	0.438 ± 0.063 a	0.51 ± 0.057 a	0.401 ± 0.062 a
		TM	0.293 ± 0.067 a	0.271 ± 0.068 a	0.169 ± 0.077 a	0.117 ± 0.065 a	0.245 ± 0.07 a
	CV2	GM	0.444 ± 0.063 a	0.477 ± 0.088 a	0.461 ± 0.111 a	0.429 ± 0.016 a	0.267 ± 0.124 a
		TM+GM	0.487 ± 0.075 a	0.504 ± 0.088 a	0.459 ± 0.115 a	0.421 ± 0.142 a	0.27 ± 0.072 a

Table 3.2: (Continued)

ISK	CV2	TM	0.109 ± 0.054 a	0.117 ± 0.079 a	0.075 ± 0.026 a	0.065 ± 0.028 a	0.008 ± 0.015 a
DON	CV1	GM	0.413 ± 0.06 a	0.273 ± 0.066 a	0.241 ± 0.072 a	0.198 ± 0.063 a	0.187 ± 0.071 a
		TM+GM	0.575 ± 0.05	0.285 ± 0.065 a	0.158 ± 0.073 a	0.226 ± 0.074 a	0.188 ± 0.07 a
		TM	0.554 ± 0.063 a	0.485 ± 0.058 ab	0.505 ± 0.064 ab	0.252 ± 0.075 b	0.469 ± 0.066 ab
INC	CV1	GM	0.56 ± 0.045 a	0.527 ± 0.052 a	0.471 ± 0.057 a	0.398 ± 0.064 a	0.406 ± 0.063 a
		TM+GM	0.525 ± 0.05 a	0.558 ± 0.051 a	0.522 ± 0.059 a	0.413 ± 0.065 a	0.411 ± 0.062 a
		TM	0.332 ± 0.075 a	0.351 ± 0.079 a	0.034 ± 0.072 b	0.08 ± 0.087 ab	0.089 ± 0.072 ab
	CV2	GM	0.426 ± 0.125 a	0.439 ± 0.088 a	0.425 ± 0.11 a	0.567 ± 0.016 a	0.297 ± 0.124 a
		TM+GM	0.394 ± 0.074 a	0.442 ± 0.071 a	0.424 ± 0.109 a	0.419 ± 0.110 a	0.296 ± 0.069 a
		TM	0.123 ± 0.091 a	0.127 ± 0.122 a	0.029 ± 0.064 a	0.012 ± 0.068 a	□0.122 ± 0.061 a
SEV	CV1	GM	0.644 ± 0.041 a	0.588 ± 0.047 a	0.614 ± 0.047 a	0.596 ± 0.049 a	0.343 ± 0.068 a
		TM+GM	0.606 ± 0.04 a	0.636 ± 0.041 a	0.52 ± 0.053 ab	0.377 ± 0.067 b	0.339 ± 0.067 b
		TM	0.381 ± 0.067 a	0.312 ± 0.066 a	0.232 ± 0.062 a	0.188 ± 0.072 a	0.244 ± 0.071 a
	CV2	GM	0.593 ± 0.054 a	0.612 ± 0.074 a	0.585 ± 0.08 a	0.338 ± 0.004 a	0.211 ± 0.071
		TM+GM	0.608 ± 0.051 a	0.624 ± 0.066 a	0.593 ± 0.085 a	0.586 ± 0.086 a	0.215 ± 0.039
		TM	0.186 ± 0.03 ab	0.227 ± 0.054 a	0.109 ± 0.013 ab	0.11 ± 0.013 ab	0.045 ± 0.030 b

†HD, days to heading; FDK, Fusarium damaged kernels; ISK, incidence, severity, and, kernel quality index; DON, deoxynivalenol; INC, incidence; SEV, severity.

‡GM, genomewide diversity array technology (DArT) markers; TM, quantitative trait loci (QTL) targeted simple sequence repeat (SSR) markers only; TM+GM, QTL targeted SSR markers and genomewide DArT markers combined.

§RF, random forest; RKHS, reproducing kernel Hilbert spaces; RR, ridge regression; BL, Bayesian Lasso; MLR, multiple linear regression.

¶Within rows, means not significantly different share a common letter

Table 3.3: Means and standard errors of fivefold cross-validation prediction accuracies for all traits using multiple linear regression models with different numbers of markers (*k*) used as fixed effects. Markers were selected based on the results of association analysis in the training set.

Trait†	Marker set‡	*k* = 5	*k* = 10	*k* = 15	*k* = 20	*k* = 25
HD	GM	0.069 ± 0.082	0.048 ± 0.086	0.07 ± 0.073	0.074 ± 0.071	0.113 ± 0.071
	TM+GM	0.071 ± 0.084	0.044 ± 0.084	0.07 ± 0.075	0.073 ± 0.071	0.107 ± 0.07
FDK	GM	0.071 ± 0.076	0.181 ± 0.063	0.313 ± 0.076	0.338 ± 0.067	0.226 ± 0.07
	TM+GM	0.067 ± 0.075	0.185 ± 0.064	0.314 ± 0.076	0.343 ± 0.068	0.222 ± 0.071
ISK	GM	-0.005 ± 0.084	0.192 ± 0.077	0.345 ± 0.072	0.4 ± 0.064	0.246 ± 0.062
	TM+GM	−0.002 ± 0.085	0.193 ± 0.078	0.348 ± 0.068	0.401 ± 0.062	0.25 ± 0.062
DON	GM	0.05 ± 0.076	0.105 ± 0.073	− 0.023 ± 0.073	0.187 ± 0.071	0.031 ± 0.066
	TM+GM	0.045 ± 0.078	0.104 ± 0.074	− 0.022 ± 0.074	0.188 ± 0.07	0.028 ± 0.068
INC	GM	0.121 ± 0.077	0.227 ± 0.074	0.33 ± 0.07	0.406 ± 0.063	0.288 ± 0.067
	TM+GM	0.116 ± 0.075	0.229 ± 0.073	0.331 ± 0.069	0.411 ± 0.062	0.291 ± 0.07
SEV	GM	0.031 ± 0.083	0.036 ± 0.08	0.283 ± 0.073	0.343 ± 0.068	0.26 ± 0.066
	TM+GM	0.029 ± 0.086	0.035 ± 0.085	0.287 ± 0.073	0.339 ± 0.067	0.267 ± 0.064

†HD, days to heading; FDK, Fusarium damaged kernels; ISK, incidence, severity, and, kernel quality index; DON, deoxynivalenol; INC, incidence; SEV, severity.‡GM, genomewide Diversity Array Technology (DArT) markers only; TM+GM, quantitative trait loci (QTL) targeted simple sequence repeat (SSR) markers and genomewide DArT markers.

Multiple linear regression accuracies obtained for these two marker sets depended largely on the trait and were highest for INC and ISK and lowest for HD. In all cases at least one of the GS models outperformed MLR. However, for some

traits, such as FDK, ISK, and INC, the difference between the mean MLR model accuracy and mean accuracy of the best GS model was surprisingly small. Furthermore, for most traits mean MLR accuracies were only slightly lower than RR accuracies. Accuracies obtained with BL were usually similar to but lower than those obtained with RR. The bivariate RR models that we evaluated for DON were as accurate as the univariate RR models. Fivefold cross-validation (CV1) accuracies for DON when using the TM+GM marker set in the bivariate models incorporating DON and either SEV, INC, or ISK were 0.219 ± 0.064, 0.24 ± 0.065, and 0.207 ± 0.062, respectively.

Nonlinear prediction model accuracies

Overall, for a given trait–marker set one of the nonparametric or semiparametric models, either RF or RKHS regression, had the highest mean accuracy in 85% of cases. Even when only QTL targeted markers were used, these models generally lead to the highest accuracies. Although RF and RKHS regression frequently led to the highest mean accuracies, in most cases they were not significantly more accurate than the other models. There was only one case where one model was significantly more accurate than all other models. When the TM+GM marker set was used to predict DON, RF regression was significantly more accurate than all other models.

Comparison of marker sets

Comparing across marker sets, accuracies resulting from the TM+GM and the GM marker sets were higher than the accuracies resulting from the TM

marker set in all cases except for HD, where all marker sets performed similarly, and DON, where the TM+GM set performed either as well or worse than the TM set depending on the model that was used (Figure 3.1).

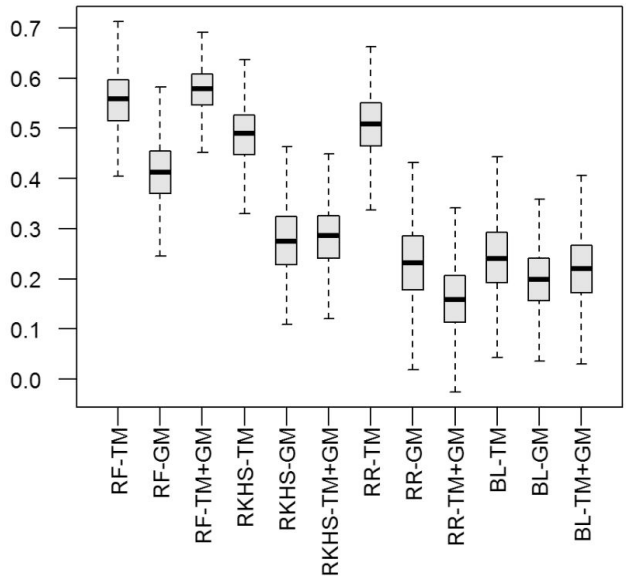


Figure 3.1: Fivefold prediction accuracies for deoxynivalenol (DON) levels using different model–marker set combinations

Prediction models include Bayesian Lasso (BL), random forest (RF) regression, reproducing kernel Hilbert spaces (RKHS) regression, and ridge regression (RR). The marker sets include genomewide Diversity Array Technology (DArT) markers (GM), Fusarium head blight (FHB) quantitative trait loci (QTL) targeted markers (TM), and both the GM and TM marker sets combined (TM+GM). For all prediction model–marker set combinations, the distribution of 1000 correlations calculated using bootstrapped samples of the 170 individuals evaluated for DON accumulation are depicted using box-and-whiskers. The black line depicts the median, the other edges of the boxes depict the lower and upper quartiles, and the outer whiskers depict the range. Outlier points are not plotted.

For DON, when using RR and RKHS regression, adding the GM marker set to the TM marker set substantially decreased the mean accuracy to the level observed when using the GM marker set alone. When using RF regression to predict DON, the TM+GM marker set led to mean accuracies equal to those observed when using the TM marker set, which was much higher than those observed when using the GM marker set alone. When using BL to predict DON, accuracies were

consistently low across all marker sets.

Assessment of unintentional prediction of maturity

Prediction models trained using PEBVs for SEV, INC, FDK, and ISK could not accurately predict the HD PEBVs of the validation set indicating that the models were not primarily capturing passive resistance due to maturity. For all models and cross-validation schemes, mean prediction accuracies for heading date were close to zero for all traits except for SEV (Table 3.4) where mean accuracies ranged between 0.02 and 0.176.

Table 3.4: Prediction accuracies for days to heading (HD) using 5 different genomic selection (GS) models trained with FHB resistance traits. Accuracies were calculated using 5-fold cross-validation (CV1) or cross-validation across years (CV2). The marker set used was a combination of both markers targeted to FHB quantitative trait loci and genome-wide diversity array technology markers (TM+GM†).

Model Training Trait‡	Cross-validation scheme	RF§	RKHS	RR	BL
FDK	CV1	-0.041 ± 0.062	-0.062 ± 0.064	0.001 ± 0.067	0.036 ± 0.065
	CV2	-0.058 ± 0.021	-0.058 ± 0.061	-0.089 ± 0.044	-0.104 ± 0.045
ISK	CV1	0.039 ± 0.063	0.05 ± 0.06	-0.047 ± 0.081	0.031 ± 0.067
	CV2	0.020 ± 0.035	0.007 ± 0.089	-0.030 ± 0.065	-0.013 ± 0.068
DON	CV1	0.072 ± 0.071	0.021 ± 0.062	-0.006 ± 0.073	-0.069 ± 0.069
INC	CV1	0.016 ± 0.059	0.006 ± 0.062	-0.078 ± 0.067	0.081 ± 0.067
	CV2	0.040 ± 0.010	0.004 ± 0.044	0.047 ± 0.019	-0.062 ± 0.03
SEV	CV1	0.151 ± 0.066	0.113 ± 0.06	0.111 ± 0.073	0.02 ± 0.068
	CV2	0.159 ± 0.073	0.176 ± 0.1	0.134 ± 0.077	0.121 ± 0.068

†TM+GM, quantitative trait loci (QTL) targeted simple sequence repeat (SSR) markers and genomewide DArT markers.
‡FDK, Fusarium damaged kernels; ISK, incidence, severity, and, kernel quality index; DON, deoxynivalenol; INC, incidence; SEV, severity.
§RF, random forest; RKHS, reproducing kernel Hilbert spaces; RR, ridge regression; BL, Bayesian Lasso.

This small positive correlation between SEV GEBVs and HD PEBVs of the validation set may indicate that the SEV prediction model incorporated some passive resistance or it could result from a genetic correlation between SEV and HD caused by pleiotrophy or linkage.

Comparison of trait-based and marker-based prediction accuracies for deoxynivalenol

Out of the four traits—ISK, SEV, INC, and FDK—ISK had the highest mean correlation with DON, $r = 0.5 \pm 0.061$, and was therefore the most predictive. The mean correlations with DON for SEV, INC, and FDK were $r = 0.432 \pm 0.073$, $r = 0.457 \pm 0.056$, and $r = 0.303 \pm 0.082$, respectively. None of these correlations are significantly different. The accuracies obtained from the methods combining ISK and markers, either the GM, TM, or TM+GM marker sets in a RF regression model are shown in Figure 2.2 where these methods are also compared to using ISK phenotypic data alone and using markers alone (either the TM or TM+GM sets) in a RF regression model. We achieved the highest mean accuracy, $r = 0.65 \pm 0.047$, from a RF regression model that used both QTL targeted markers and the correlated trait ISK as the predictor variables; the next best models were (i) RF regression models incorporating QTL targeted, genomewide DArT markers and ISK, $r = 0.616 \pm 0.049$, and (ii) incorporating only genomewide DArT markers and ISK, $r = 0.527 \pm 0.054$.

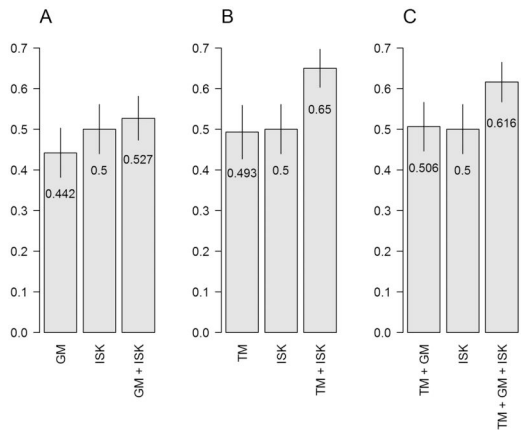


Figure 3.2: Comparison of mean fivefold cross-validation prediction accuracies for deoxynivalenol (DON) levels using only markers in a random forest (RF) regression model, only incidence, severity, and, kernel quality index (ISK) phenotypic values, or ISK values and markers combined in a RF regression model.

The RF regression models based on marker information include A) genomewide Diversity Array Technology (DArT) markers (GM), B) Fusarium head blight (FHB) quantitative trait loci (QTL) targeted markers (TM), and C) models trained using both GM and TM targeted markers (TM+GM). For each of these marker sets, RF regression accuracies for DON are compared with accuracies obtained using only correlation with ISK phenotypic values and accuracies obtained from using both markers and ISK values in a RF regression model. These combination models include A) GM+ISK, B) TM+ISK, and C) TM+GM+ISK. For each prediction method, error bars depict standard errors.

Discussion

Prediction strategies

It appeared that FHB resistance traits fell into two distinct categories based on the prediction strategies that led to the highest accuracies. The first category included SEV, INC, ISK, and FDK. The prediction strategies leading to the highest accuracies for these traits were characteristic of a GS approach and are what we would expect for most quantitative traits. Specifically, (i) GS models outperformed MLR models, (ii) predictions based on QTL targeted markers alone were low, and (iii) adding QTL targeted markers in addition to genomewide markers did not improve accuracy. The second category included only DON. With

DON the prediction strategies that led to the highest accuracies were more characteristic of a marker-assisted recurrent selection approach. Specifically, (i) MLR models and GS models lead to similar accuracies, (ii) predictions based on QTL targeted markers alone were higher than predictions based on both QTL targeted and genomewide markers, and (iii) RF, which appeared to better ignore uninformative predictors, was significantly more accurate than other prediction models when both QTL targeted and genomewide markers were used. Based on these trends it appears that fewer loci are involved in DON resistance compared to the other traits and a GS model using targeted QTL only appears to be the appropriate approach for prediction.

Breeding strategies for deoxynivalenol using correlated trait information

Because we found that predictions based on a RF regression model incorporating markers were as accurate as those based only on phenotypes for ISK, it would be possible to realize the same genetic gain per cycle for DON with a GS model as with selection using ISK as a proxy for DON in phenotypic selection. The advantage to using a GS model is that phenotyping the correlated traits is not required before selection, which enables more cycles of selection to be achieved per unit of time. Therefore, selection for DON resistance based on a GS model rather than on correlated traits could lead to greater genetic gain per unit of time because selection can occur in a greenhouse or off-season nursery.

Although selecting for DON based on a GS model would lead to greater gain per unit of time by way of accelerating the breeding process, for all the

prediction methods that we compared, the highest gain per cycle was predicted to be achieved by incorporating ISK and markers into a RF regression model. The disadvantage to this prediction method is that phenotypic data for ISK must be available. Therefore, unless genotyping for important QTL and evaluating ISK is less costly than evaluating DON directly, there may be no benefit to using this type of model.

Prediction model performance

The consistently better performance of RF and RKHS regression across different traits and marker sets may be due to the design of the cooperative nursery and/or the genetic architecture of the traits. The design of the cooperative nurseries was to use relatively few lines evaluated many times and this may lead to higher accuracies for RF and RKHS regression relative to BL, RR, and MLR. This is because RF and RKHS models do not estimate marker effects; instead, they predict genetic values of unobserved lines based on genetic similarity to lines in the training set. Therefore, greater replication leading to increased accuracy in the genetic value estimations of the lines themselves may provide greater benefit to RF and RKHS regression compared to RR, BL, and MLR. A more powerful design for marker effect estimation would consist of a large number of lines with minimal replication (Knapp and Bridges, 1990).

Alternatively, RF and RKHS regression could be more predictive because the loci underlying the variation for the traits we analyzed do not behave strictly additively. Random forest and RKHS regression are able to capture these

nonadditive effects and therefore may be more accurate than RR, BL, and MLR when nonadditive effects are important. This possible explanation is supported by previous studies that have documented the nonadditive behavior of FHB resistance loci (Yang et al., 2005; Ma et al., 2006; Guo et al., 2006; Pumphrey et al., 2007; Yu et al., 2007).

Conclusions

This study found that data from the U.S. cooperative FHB nurseries, which consist of germplasm from many different institutions evaluated in different regions for FHB resistance, can be used to train relatively accurate prediction models that could be useful in breeding for native FHB resistance. For DON resistance breeding specifically, we found that selection based on marker based prediction models could lead to greater genetic gain per cycle and greater genetic gain per unit time compared to selection based on correlated traits alone.

This study also shows that prior information about DON can be used to improve genomic prediction accuracies. Therefore, GS models incorporating such information should be evaluated for traits of interest if possible. Prior information about QTL can be used to ensure that the markers used for prediction are linked to specific QTL of known importance. This is especially important if the random genomewide markers used for prediction are in low LD with the predictive QTL regions. Although in theory genomewide markers should be adequate to capture QTL effects, in practice the markers may not be distributed randomly across the genome and may be completely missing some

segregating segments. In addition, prediction strategies based only on markers linked to important QTL should be evaluated because for certain traits they may be more useful for prediction and genomewide markers may only capture noise. In addition to incorporating prior information about important QTL, this study shows that incorporating information about correlated phenotypes can be beneficial. If data on a correlated trait is available, prediction models incorporating that phenotypic data should be evaluated.

Although this work demonstrates that GS can be successful for cases such as FHB resistance in U.S. wheat germplasm, it also points out issues that should be further studied. Most importantly, studies aiming to empirically evaluate breeding strategies that implement GS need to be conducted and compared to conventional strategies. Second, studies of the implications of using nonadditive models such as RF and RKHS regression across multiple cycles of selection are needed before such methods can be recommended. Although RF and RKHS regression often lead to high accuracies it is not clear how much of the additive genetic variation such models are capturing relative to nonadditive genetic information and in a recurrent selection context, predictions based partially on nonadditive effects could lead to lower gain from selection than expected. In addition, continued research to evaluate the benefit of marker selection or the targeted genotyping of specific loci for use in GS is warranted to determine under what circumstances such an approach may be useful. Lastly, further work to develop GS models that can incorporate both marker and phenotypic

information, such as the models we evaluated for DON resistance, may be useful for improving prediction accuracies.

Acknowledgments

This research was supported in part by the United States Department of Agriculture[1]-Agricultural Research Service (USDA-ARS) -NIFA-AFRI grants, award numbers 2009-65300-05661, 2011-68002-30029, and 2005-05130, and by Hatch project 149-449. In addition, partial support for J. Rutkoski was provided by a USDA National Needs Fellowship Grant #2008-38420-04755.

References

Akbari, M., P. Wenzl, V. Caig, J. Carling, L. Xia, S.Y. Yang, G. Uszynski, V. Mohler, A. Lehmensiek, H. Kuchel, M.J. Hayden, N. Howes, P. Sharp, P. Vaughan, B. Rathmell, E. Huttner, and A. Kilian. 2006. Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. Theor. Appl. Genet. 113: 1409–1420.

Anderson, J.A., S. Chao, and S. Liu. 2007. Molecular breeding using a major QTL for Fusarium head blight resistance in wheat. Crop Sci. 47(S3): S112–S119.

Asoro, F.G., M.A. Newell, W.D. Beavis, M.P. Scott, and J.-L. Jannink. 2011. Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. Plant Gen. 4: 132–144.

Bates, D., M. Maechler, and B. Dai. 2009. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-39. Institute for Statistics and Mathematics, Wien, Austria. http://cran.r-project.org/web/packages/lme4 (accessed 12 May 2010).

Benson, J., and G. Brown-Guedira. 2012. Population structure and linkage disequilibrium of winter wheat in regional Fusarium head blight screen nurseries. Plant Gen. 5: 71-80.

Bernardo, A.N., H. Ma, D. Zhang, and G. Bai. 2011. Single nucleotide polymorphism in wheat chromosome region harboring *Fhb1* for Fusarium head blight resistance. Mol. Breed. 29: 477–488.

L. Breiman 2001. Random Forests. Mach. Learn. 45: 5–32.

Buerstmayr, H., T. Ban, and J.A. Anderson. 2009. QTL mapping and marker-assisted selection for Fusarium head blight resistance in wheat: A review. Plant Breed. 128: 1–26.

Buerstmayr, H., M. Lemmens, L. Hartl, L. Doldi, B. Steiner, M. Stierschneider, and P. Ruckenbauer. 2002. Molecular mapping of QTLs for Fusarium head blight resistance in spring wheat. I. Resistance to fungal spread (Type II resistance). Theor. Appl. Genet. 104: 84–91.

Buerstmayr, H., M. Lemmens, M. Schmolke, G. Zimmermann, L. Hartl, F. Mascher, M. Trottet, N.E. Gosman, and P. Nicholson. 2008. Multi-environment evaluation of level and stability of FHB resistance among parental lines and selected offspring derived from several European winter wheat mapping populations. Plant Breed. 127: 325–332.

Carlson, C.S., M.A. Eberle, M.J. Rieder, Q. Yi, L. Kruglyak, and D.A. Nickerson. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am. J. Hum. Genet. 74: 106–120.

Casale, W.L., J.J. Pestka, and L.P. Hart. 1988. Enzyme-linked immunosorbent assay employing monoclonal antibody specific for deoxynivalenol (vomitoxin) and several analogs. J. Agric. Food Chem. 36: 663–668

Chen, J., C.A. Griffey, M.A.S. Maroof, E.L. Stromberg, R.M. Biyashev, W. Zhao, M.R. Chappell, T.H. Pridgen, Y. Dong, and Z. Zeng. 2006. Short communication: Validation of two major quantitative trait loci for Fusarium head blight resistance in Chinese wheat line W14. Plant Breed. 101: 2003–2005.

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueño, J.L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger, and H.-J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713–724.

Cuthbert, P.A., D.J. Somers, and A. Brulé-Babel. 2007. Mapping of *Fhb2* on chromosome 6BS: A gene controlling Fusarium head blight field resistance in bread wheat (*Triticum aestivum* L.). Theor. Appl. Genet. 114: 429–437.

de los Campos, G., D. Gianola, and G.J.M. Rosa. 2009a. Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation. J. Anim. Sci. 87: 1883–1887.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J.M. Cotes. 2009b. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182: 375–385.

de los Campos, G., and P. Perez Rodriguez. 2010. BLR: Bayesian linear regression. R package version 1.2. Institute for Statistics and Mathematics, Wien, Austria. http://cran.r-project.org/web/packages/BLR/index.html (accessed 26 Apr. 2011).

B. Efron 1979. Bootstrap methods: Another look at the jackknife. Ann. Stat. 7: 1–26.

Efron, B., and R. Tibshirani. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat. Sci. 1: 54–75.

Garrick, D.J., J.F. Taylor, and R.L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41:55.

Gianola, D., R.L. Fernando, and A. Stella. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173: 1761–1776.

Gilbert, J., and S.M. Woods. 2006. Strategies and considerations for multi-location FHB screening nurseries. In: Ban, T., Lewis, J.M., and Phipps, E.E., editors, The Global Fusarium Initiative for International Collaboration: A Strategic Planning Workshop, CIMMYT, El Batan, Mexico. 14–17 Mar. 2006. CIMMYT, El Batan, Mexico. p. 93–102.

Gilmour, A.R., B.J. Gogel, B.R. Cullis, and R. Thompson. 2009. ASReml user guide release 3.0. VSN International Ltd, Hemel Hempstead, UK.

González-Recio, O., and S. Forni. 2011. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genet. Sel. Evol. 43: 7.

Gosman, N., R. Bayles, P. Jennings, J. Kirby, and P. Nicholson. 2007. Evaluation and characterization of resistance to Fusarium head blight caused by Fusarium culmorum in UK winter wheat cultivars. Plant Pathol. 56: 264–276.

Guo, P.-G., G.-H. Bai, R.-H. Li, G. Shaner, and M. Baum. 2006. Resistance gene analogs associated with Fusarium head blight resistance in wheat. Euphytica 151: 251–261.

Heffner, E.L., J.-L. Jannink, H. Iwata, E. Souza, and M.E. Sorrells. 2011a. Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Sci. 51: 2597–2606.

Heffner, E.L., J.-L. Jannink, and M.E. Sorrells. 2011b. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Gen. 4: 65–75.

Heffner, E.L., A.J. Lorenz, J.-L. Jannink, and M.E. Sorrells. 2010. Plant breeding with genomic selection: Gain per unit time and cost. Crop Sci. 50: 1681-1690.

Heslot, N., H.-P. Yang, M.E. Sorrells, and J.-L. Jannink. 2012. Genomic selection in plant breeding: A comparison of models. Crop Sci. 52: 146–160.

Jiang, G.-L., Y. Dong, J. Shi, and R.W. Ward. 2007a. QTL analysis of resistance to Fusarium head blight in the novel wheat germplasm CJ 9306. II. Resistance to deoxynivalenol accumulation and grain yield loss. Theor. Appl. Genet. 115: 1043–1052.

Jiang, G.-L., J. Shi, and R.W. Ward. 2007b. QTL analysis of resistance to Fusarium head blight in the novel wheat germplasm CJ 9306. I. Resistance to fungal spread. Theor. Appl. Genet. 116: 3–13.

Kang, H.M., N.A. Zaitlen, C.M. Wade, A. Kirby, D. Heckerman, M.J. Daly, and E. Eskin. 2008. Efficient control of population structure in model organism association mapping. Genetics 178: 1709–1723.

Knapp, S.J., and W.C. Bridges. 1990. Using molecular markers to estimate quantitative trait locus parameters: Power and genetic variances for unreplicated and replicated progeny. Genetics 126: 769–777.

Kolb, F.L., and L.K. Boze. 2003. An alternative to the FHB index: incidence, severity, kernel rating (ISK) index. In: Cantry, S.M., Lewis, J., and Ward, R.W. , editors, Proceedings of the National Fusarium Head Blight Forum, Bloomington, MN. 13–15 Dec. 2003. Michigan State University, East Lansing, MN. p. 259.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R News 2: 18–22.

Liu, S., M.D. Hall, C.A. Griffey, and A.L. McKendry. 2009. Meta-analysis of QTL associated with Fusarium head blight resistance in wheat. Crop Sci. 49: 1955–1968.

Lorenz, A.J., S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata, K.P. Smith, M.E. Sorrells, and J.-L. Jannink. 2011. Genomic selection in plant breeding: Knowledge and prospects. Adv. Agron. 110: 77–123.

Lorenz, A.J., K.P. Smith, and J.-L. Jannink. 2012. Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Sci. 52: 1609-1621.

Lorenzana, R.E., and R. Bernardo. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. 120: 151–161.

Ma, H.-X., G.-H. Bai, X. Zhang, and W.-Z. Lu. 2006. Main effects, epistasis, and environmental interactions of quantitative trait loci for Fusarium head blight resistance in a recombinant inbred population. Phytopathology 96: 534–541.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Miedaner, T., C. Reinbrecht, U. Lauber, M. Schollenberger, and H. Geiger. 2002. Effects of genotype and genotype-environment interaction on deoxynivalenol accumulation and resistance to Fusarium head blight in rye, triticale, and wheat. Plant Breed. 120: 97–105.

Miedaner, T., T. Würschum, H.P. Maurer, V. Korzun, E. Ebmeyer, and J.C. Reif. 2010. Association mapping for Fusarium head blight resistance in European soft winter wheat. Mol. Breed. 28: 647–655.

Nganje, W.E., D.A. Bangsund, F.L. Leistritz, W.W. Wilson, and N.M. Tiapo. 2004. Regional economic impacts of Fusarium head blight in wheat and barley. Rev. Agric. Econ. 26:332–347.

Pallotta, M.A., P. Warner, R.L. Fox, H. Kuchel, S.J. Jefferies, and P. Langridge. 2003. Marker assisted wheat breeding in the southern region of Australia. In: Pogna, N.E., editor, Proceedings of the 10th International Wheat Genetics Symposium, Paestum, Italy. 1–6 Sept. 2003. Istituto Sperimentale per la Cerealicoltura, Rome, Italy. p. 789–791.

Park, T., and G. Casella. 2008. The Bayesian Lasso. J. Am. Stat. Assoc. 103: 681–686.

Pathre, S.V., and C.J. Mirocha. 1977. Assay methods for trichothecenes and review

of their natural occurrence. In: Rodricks, J.V., Hesseltine, C.W., and Mehlman, M.A., editors, Mycotoxins in human and animal health. Pathotox Publishers, Park Forest South, IL. p. 229–253.

Paul, P.A., P.E. Lipps, and L.V. Madden. 2005. Relationship between visual estimates of Fusarium head blight intensity and deoxynivalenol accumulation in harvested wheat grain: A meta-analysis. Phytopathology 95: 1225–1236.

Pérez, P., G. de los Campos, J. Crossa, and D. Gianola. 2010. Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. Plant Gen. 3: 106–116.

Pestka, J.J., and A.T. Smolinski. 2005. Deoxynivalenol: Toxicology and potential effects on humans. J. Toxicol. Environ. Health B 8: 39–69.

H.P. Piepho 2009. Ridge regression and extensions for genomewide selection in maize. Crop Sci. 49: 1165–1176.

Pumphrey, M.O., R. Bernardo, and J.A. Anderson. 2007. Validating the QTL for Fusarium head blight resistance in near-isogenic wheat lines developed from breeding populations. Crop Sci. 47: 200–206.

R Development Core Team. 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org (accessed 29 June 2011).

SAS Institute. 2010. JMP Genomics user guide. SAS Institute, Cary, NC.

Sneller, C., M. Guttieri, P. Paul, J. Costa, and R. Jackwood. 2012. Variation for resistance to kernel infection and toxin accumulation in winter wheat infected with Fusarium graminearum. Phytopathology 102:306–314.

Sneller, C.H., P. Paul, and M. Guttieri. 2010. Characterization of resistance to Fusarium head blight in an eastern U.S. soft red winter wheat population. Crop Sci. 50: 123–133.

Somers, D.J., G. Fedak, and M. Savard. 2003. Molecular mapping of novel genes controlling Fusarium head blight resistance and deoxynivalenol accumulation in spring wheat. Genome 564: 555–564. doi:10.1139/g03-033

Whittaker, J.C., R. Thompson, and M.C. Denham. 2000. Marker-assisted selection using ridge regression. Genet. Res. 75: 249–252.

C.E. Windels 2000.Economic and social impacts of Fusarium head blight: Changing farms and rural communities in the northern great plains. Phytopathology 90: 17–21.

Yang, Z., J. Gilbert, G. Fedak, and D.J. Somers. 2005. Genetic characterization of QTL associated with resistance to Fusarium head blight in a doubled-haploid spring wheat population. Genome 48: 187–196.

Yu, J.-B., G.-H. Bai, W.-C. Zhou, Y.-H. Dong, and F.L. Kolb. 2007. Quantitative trait loci for Fusarium head blight resistance in a recombinant inbred population of Wangshuibai/Wheaton. Phytopathology 98: 87–94.

Yu, J.M., G. Pressoir, W.H. Briggs, I.V. Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38: 203–208.

Zhou, W., F.L. Kolb, G. Bai, G. Shaner, and L.L. Domier. 2002. Genetic analysis of scab resistance QTL in wheat with microsatellite and AFLP markers. Genome 727: 719–727.

CHAPTER 4

GENOMIC SELECTION FOR QUANTITATIVE ADULT PLANT STEM RUST RESISTANCE IN WHEAT[4]

Abstract

Quantitative adult plant resistance (APR) to stem rust (*Puccinia graminis* f. sp. *tritici*) is an important breeding target in wheat (*Triticum aestivum* L.) and a potential target for genomic selection (GS). To evaluate the relative importance of known APR loci in applying genomic selection, we characterized a set of CIMMYT germplasm at important APR loci and on a genome-wide profile using genotyping-by-sequencing. Using this germplasm, we describe the genetic architecture and evaluate prediction models for APR using data from the international Ug99 stem rust screening nurseries. Prediction models incorporating markers linked to important APR loci and seedling phenotype scores as fixed effects were evaluated along with the classic prediction models: Multiple linear regression (MLR), Genomic best linear unbiased prediction (G-BLUP), Bayesian Lasso (BL), and Bayes Cπ (BCπ). We found the *Sr2* region to play an important role in APR in this germplasm. A model using *Sr2* linked markers as

[4] Originally published as: Rutkoski J. E., J. Poland, R.P. Singh, J. Huerta-Espino, S. Bhavani, M. Rouse, H. Barbier, J-L. Jannink, M. E. Sorrells. Genomic selection for quantitative adult plant stem rust resistance in wheat. Plant Gen. (in press).

fixed effects in G-BLUP was more accurate than MLR with *Sr2* linked markers (p-value = 0.12), and ordinary G-BLUP (p-value = 0.15). Incorporating seedling phenotype information as fixed effects in G-BLUP did not consistently increase accuracy. Overall, levels of prediction accuracy found in this study indicate that GS can be effectively applied to improve stem rust APR in this germplasm, and if genotypes at *Sr2* linked markers are available, modeling these genotypes as fixed effects could lead to better predictions.

Abbreviations

APR, Adult plant resistance; BCπ, Bayes Cπ; BL, Bayesian Lasso; G-BLUP, Genomic best linear unbiased prediction; GS, Genomic selection; MLR, Multiple linear regression; QTL, Quantitative trait loci; STS, Sequence tagged site.

Introduction

Stem rust, caused by *Puccinia graminis* f. sp. *tritici*, is a globally widespread and highly damaging disease of wheat (*Triticum aestivum* L.) capable of causing up to 100% yield losses in susceptible varieties (Park, 2007). After adoption of resistant varieties during the 1950s, outbreaks of stem rust became rare. However, the recent emergence of a new stem rust race group named Ug99 (Pretorius et al., 2000) capable of infecting the majority of the worlds' wheat germplasm (Singh et al., 2006), has highlighted the need for breeding efforts focused on durable stem rust resistance.

Resistance to stem rust generally falls into two categories: 1) All stage resistance which is often conferred by race-specific genes involved in pathogen

recognition and associated with a hypersensitive response, and 2) Slow rusting adult plant resistance (APR) which is quantitative resistance often conferred by multiple loci, and is not associated with a hypersensitive response. Quantitative resistance is usually considered more durable than that conferred by pathogen recognition genes (Parlevliet, 2002), however, it must be improved over multiple cycles of selection using well managed screening nurseries for evaluation.

Genomic selection (GS) (Meuwissen, Hayes, & Goddard, 2001), reviewed by Lorenz et al. (2011) and Heffner et al. (2009) is breeding technology that may increase rates of genetic gain for quantitative traits. With GS, a genomic prediction model is used to predict breeding values of selection candidates, and selections are made based on these predictions. A model training population consisting of relevant individuals that have been both genotyped and phenotyped is used to calibrate the prediction model.

Various genomic prediction models have been developed. Models differ according to how markers of different effect sizes are treated. Genomic best linear unbiased prediction (G-BLUP, Bernardo, 1994; Piepho, 2009), treats markers homogenously, whereas Bayesian methods such as Bayesian Lasso (BL, Park and Casella, 2008) and Bayes Cπ (BCπ, Habier et al., 2011) treat markers of different effect sizes heterogeneously. Such methods are expected to better model traits with large-effect quantitative trait loci (QTL).

Because moderate effect genes, such as *Sr2*, and *Lr34,* also known as *Sr57*, are known to be involved in stem rust APR (Sunderwirth and Roelfs, 1980; Dyck,

1987; Singh et al., 2012), prediction models that attempt to realistically model these loci may be more accurate than a standard G-BLUP model. Markers linked to these loci could be predictive alone or modeled as fixed effects in combination with genome-wide markers. Similarly, seedling resistance phenotypes, which are often collected in addition to APR, could be useful fixed effects predictor variables. The objective of this study was to compare prediction models for stem rust APR and to determine if explicitly modeling large-effect loci or seedling phenotypes as fixed effects in a G-BLUP model could lead to higher accuracies than those achieved with G-BLUP or Bayesian models.

Materials and methods

Phenotypic data

Adult plant stage*:* Three hundred sixty five advanced CIMMYT breeding lines were used in all analyses. Quantitative stem rust APR was phenotyped at the international Ug99 stem rust screening nurseries: Kenya Agricultural Research Institute, Njoro, Kenya and the Ethiopian Institute of Agricultural Research, Debre Zeit, Ethiopia between 2007 and 2012 as described in Yu et al. (2011). Data was from 12 environments (location/season combinations), three of which were at Debre Zeit. 'Kingbird' and 'PBW343' served as moderately resistant and moderately susceptible check varieties. Each breeding line, excluding the checks, appeared in approximately four of the 12 environments, and appeared only once per environment. Each plot consisted of two 70cm rows spaced 30 cm apart. Disease severity was measured visually on a modified Cobb scale (Peterson et al.,

1948). Measuremets were taken between the early and late dough stage and a week to ten days later. Phenotypic distributions within environments are shown in Figure 4.1. A Box-Cox transformation was applied prior to all analyses (Box and Cox, 1964) to avoid non-normal residuals.

Seedling stage: Lines were evaluated at the seedling stage for reaction to Ug99 stem rust race TTKSK, isolate 04KEN156/04, at the USDA-ARS Cereal Disease Laboratory using cool and normal post-inoculation temperature treatments. Seedlings were inoculated as in Jin et al. (2007) and then placed in a growth chamber with a 14 hour photoperiod at 18°C day and 15°C night for the cool treatment and 22°C day and 19°C night for the normal treatment. Seedling evaluations at both cool and normal treatments were replicated twice. Infection types on a zero to four scale as in Stakman et al. (1962) were recorded 14 days post-inoculation and then converted into a numerical value from zero to nine as described by Zhang et al. (2011). Stakman infection types greater than or equal to '3' were considered high infection types. Infection type ';' describes the observation of visible chlorotic spots associated with hypersensitive resistance. When multiple infection types were observed on a single leaf, all infection types were recorded starting with the most commonly observed infection type.

Heritability estimation

Broad sense heritability (H^2) on a line mean basis was calculated according to Hallauer et al., (2010). Variance components were estimated in R version 3.0.1 (R Development Core Team, 2010) using the package *lme4* (Bates

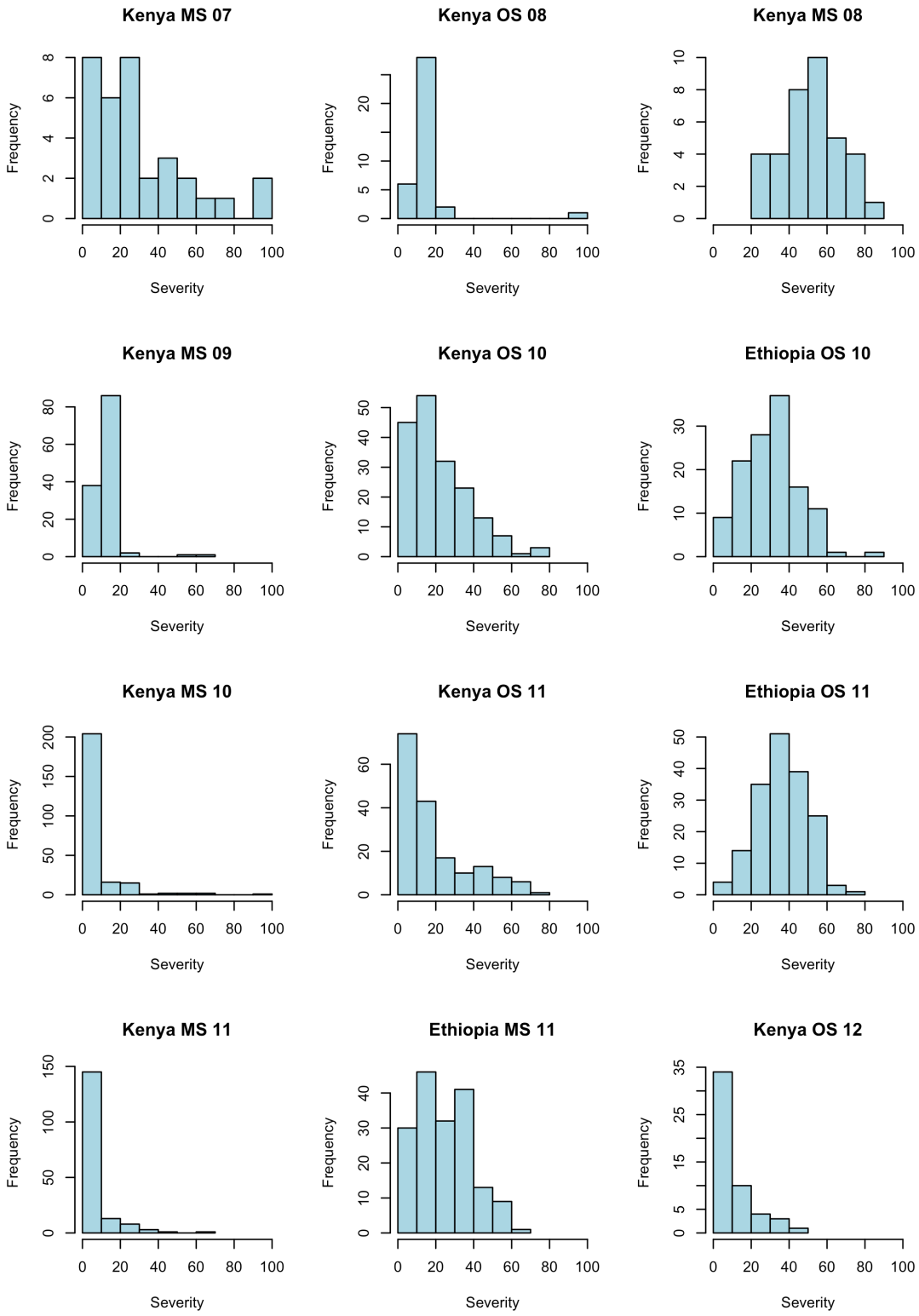


Figure 4.1: Phenotypic distributions of stem rust severity within each environment OS: off-season, MS: main-season

and Maechler, 2010).

Genotypic data

Genome-wide-genotyping: Genotyping-by-sequencing (GBS, Elshire et al., 2011) was used to generate genome-wide markers according to the protocol described in Poland et al. (2012a). A total of 27,434 polymorphisms were detected. Missing data were imputed using random forest imputation described in Poland et al. (2012b) as recommended by Rutkoski et al. (2013). Markers with greater than 50% missing data were removed and a set of non-redundant GBS markers with pairwise r^2 values less than 0.8 were selected (Carlson et al., 2004) leaving 4040 markers.

Loci targeted genotyping: Markers targeted to *Sr2* and *Lr34*, were genotyped using sequence tagged site (STS), simple sequence repeat and KASPar™ (www.lgcgenomics.com) assays. All KASPar™ assays were run at the Eastern Regional Small Grains Genotyping Laboratory, Raleigh, North Carolina. For *Lr34*, two gene based KASPar™ assays were used to determine presence or absence of the resistance allele based on sequence polymorphism reported by Lagudah et al., (2009). The STS marker *csLV34* (Lagudah et al., 2006), 0.4 cM from *Lr34*, was also assayed. For *Sr2*, the simple sequence repeat marker *gwm*533 (Spielmeyer et al., 2003), the STS marker *csSr2* (Mago et al., 2011) and a KASPar™ assay based on the polymorphism targeted by *csSr2* (referred to as *csSr2_KASPar*) were used.

Genotypic value estimation

The R package *rrBLUP* (Endelman, 2011) was used to calculate the restricted maximum likelihood (REML) solutions for the mixed model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

where **Y** was the vector of phenotypes, **β** was the vector of environment effects treated as fixed, **u** was the vector of genotype effects treated as random, **X** and **Z** were the design matrices relating **β** and **u** to the observations in **Y** and **ε** is the residual error. Genetic values, **u** were de-regressed according to Garrick et al. (2009). De-regressed genetic values, $\mathbf{Y_{GV}}$, were calculated as

$$\mathbf{Y_{GV}} = \frac{\mathbf{u}}{\mathbf{1} - \frac{\mathbf{PEV}}{\sigma_u^2}}$$

where is the genetic variance, and PEV is the prediction error variance. Solutions for both σ_u^2 and **PEV** were returned from the mixed model fit using *rrBLUP*. De-regressed genetic values, $\mathbf{Y_{GV}}$, were used to validate prediction models. De-regression was appropriate because individuals had different numbers of observations. Genetic values for individuals with few observations are shrunk more towards zero than genetic values of individuals with many observations.

Genome wide association

Genome wide association was performed using a mixed model accounting for kinship (Yu et al., 2006). According to Kang et al. (2010), variance components were estimated once by fitting the mixed model:

$$\mathbf{Y_{GV}} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}$$

$\mathrm{Var}(\mathbf{u}) = \mathbf{G}\sigma_u^2$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{I}\sigma_u^2$. **G** is a marker relationship matrix which was calculated according to VanRaden (2008) implemented in the R package GAPIT (Lipka et al., 2012). For each marker *k* with MAF≥0.05, a total of 3903 markers, we estimated its effect $\boldsymbol{\beta}_k$ and F-statistic, testing the null hypothesis that $\boldsymbol{\beta}_k$=0 in the model:

$$\mathbf{Y_{GV}} = \mathbf{X\beta} + \mathbf{X}_k\boldsymbol{\beta}_k + \boldsymbol{\eta}$$

$\boldsymbol{\beta}_k$ is the effect of marker *k*, $\mathbf{X}_k$ is the genotype of marker *k*, and $\hat{\sigma}_{\boldsymbol{\eta}}^2 = \hat{\sigma}_u^2\mathbf{ZGZ}' + \hat{\sigma}_{\boldsymbol{\varepsilon}}^2\mathbf{I}$. One thousand permutations (Churchill and Doerge, 1994) were used to calculate the p-value significance threshold at an experimentwise α of 0.05.

Prediction models

Fixed effects models: Two multiple linear regression (MLR) methods were used, A and B. MLR A consisted of a marker selection and marker effect estimation step. Both marker selection and marker estimation were carried out within the model training set only. For variable selection, p-values from a genome-wide association analysis were used to rank markers. No kinship correction was used because markers that capture kinship are useful for prediction within the population of interest, even though they may not be linked to causative loci. Then, for each iteration *i* through *l,* a marker was added to the model:

$$\mathbf{Y_{GV}} = \mathbf{1}\boldsymbol{\beta}_0 + \mathbf{X}_i\boldsymbol{\beta}_i \ldots \mathbf{X}_l\boldsymbol{\beta}_l + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\beta}_0$ is the mean, $\boldsymbol{\beta}_i$ is the effect of marker *i*, and $\mathbf{X}_i$ is the genotype of marker *i*. After each iteration, the 5-fold cross validation accuracy was calculated within the training set and when accuracy$_{l-1}$ > accuracy$_l$ the model with *l*-1 markers was selected. Predicted breeding values of an individual *j,* were calculated as:

$$\hat{y}_j = \hat{\beta}_0 + \sum_i^{i-1} \hat{\beta}_i x_{ij}$$

For MLR B, the marker selection step was done only among the five markers linked to candidate genes.

Mixed models: For G-BLUP (Bernardo, 1994; Piepho, 2009), breeding values were predicted using the mixed model.

$$\mathbf{Y}_{\mathrm{GV}} = \mathbf{1}_n \boldsymbol{\beta}_0 + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

$$\mathbf{u} \sim \mathrm{N}(0, \mathbf{G}\hat{\sigma}_{\mathrm{u}}^2)$$

where the solutions for **u** consist of the genomic estimated breeding values. G-BLUP was implemented using the R package *rrBLUP* (Endelman, 2011). G-BLUP A was a version of G-BLUP that included selected markers as fixed effects in the G-BLUP model and all markers as random effects. By selecting markers as fixed effects, we assume that each selected marker has a unique variance. For fixed effect variable selection, p-values from a genome-wide association analysis without structure correction were used to rank markers, then for each iteration *i* through *l,* a marker was added to the model:

$$\mathbf{Y}_{\mathbf{GV}} = \mathbf{1}\boldsymbol{\beta}_0 + \mathbf{X}_i\boldsymbol{\beta}_i \ldots \mathbf{X}_l\boldsymbol{\beta}_l + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

$$\mathbf{u} \sim \mathrm{N}(0, \mathbf{G}\hat{\sigma}_{\mathrm{u}}^2)$$

for each iteration 5-fold cross validation accuracy within the training set was calculated. When accuracy$_{l-1}$ > accuracy$_{l}$, the model with l-1 fixed effect markers was selected. Predicted breeding values of each individual j, were calculated as:

$$\hat{y}_j = \hat{\beta}_0 + \sum_i^{i-1} \hat{\beta}_i x_{ij} + u_j$$

For G-BLUP B the fixed effect marker selection step was done only among the five markers linked to candidate genes. For G-BLUP T the fixed effects were the seedling phenotypes for the normal and cool treatments.

Bayesian models: The general model for BL (Park and Casella, 2008) and BCπ was:

$$\mathbf{Y_{GV}} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

X is a design matrix for the markers, and **β** is a vector of m marker effects. Predicted breeding values were estimated as:

$$\hat{\mathrm{y}}_j = \hat{\beta}_0 + \sum_i^m \hat{\beta}_i x_{ij}$$

For BL the marginal prior of marker effects was double exponential (Pérez et al., 2010). BL was implemented in the R package *BLR* (de los Campos and Perez Rodriguez, 2010). For BCπ (Habier et al., 2011) the prior for $\boldsymbol{\beta}_i$ depends on a common marker variance and the prior probability, π, that marker i has no effect. The priors and prior parameters were as described in {Habier et al., (2011). BCπ was implemented in R using code adapted from R.L. Fernando. For both BL and BCπ a total of 60,000 iterations were used and the first 20,000 were excluded as burn-in.

Prediction model accuracy calculation

Prediction accuracies were calculated using 10-fold cross validation. Cross validation folds were selected to be representative samples using cluster assignment information from hierarchical agglomerative clustering (Fraley and Raftery, 2002) implemented using the R package *mclust* (Fraley et al., 2012). One accuracy value was computed for each model by computing the Pearson's correlation (r) between the de-regressed genetic values $\mathbf{Y_{GV}}$ and the predicted breeding values. Accuracies were computed using two different marker sets: GBS markers only, and all available markers. In addition to accuracy, Spearman's rank correlations between the estimated breeding values for all possible pairs of prediction models was computed to compare prediction model outcomes.

Significance testing among prediction model accuracies

Statistical significance between prediction model accuracies was determined using paired, two-sided t-tests carried out by bootstrapping. The inference space for model comparison was CIMMYT spring wheat absent of major genes effective against stem rust race TTKST, evaluated for stem rust adult plant resistance between 2007 and 2012, and identified as candidates for release to international partners, a population of about 500 lines. The set of 365 individuals from that population was randomly split into a training set of 265 individuals and a validation set of 100 individuals. Then, for each iteration bootstrapped samples of the training set and validation sets were drawn. To simulate the sampling variability of polymorphisms detected using GBS, a sample of GBS markers of size

2694 (2/3 of the total markers) was also drawn. This is equivalent to taking a bootstrap sample of markers and then only using non-redundant markers for model fitting. Selection of non-redundant markers is a common practice prior to GWAS or GS. Using this sampled dataset, prediction accuracy was measured using all prediction models except BCπ and BL, which were excluded to reduce computational burden. This process was repeated for 1000 iterations. For a given pair of models, the accuracy vectors were subtracted to create a distribution of differences. A two-tailed p-value was calculated by calculating the frequency of values above or below 0, multiplied by two. Mean accuracies for each model were also calculated.

The bootstrap t-testing procedure for model comparison relies on several assumptions. The first assumption is that the sample of 275 individuals in the training set and the sample of 100 individuals in the validation set are representative of the population from which they were originally sampled, which is met as long as the samples are sufficiently large and selected from the population at random. The second assumption is that the observations, in our case de-regressed genetic values, are independent. Non-independence can arise if the values consist of repeated measurements on the same individuals or if the data consists of clusters of individuals more similar to each other than what would be expected based on random sampling from the original population. The third assumption is that the observations are identically distributed, meaning that there are no systematic trends in the mean or variance of the values.

Results

Phenotypic data

Adult plant stem rust resistance was highly heritable, with a line mean broad sense heritability of 0.82. The absence of race-specific resistance genes effective against TTKST in the set of 365 lines was confirmed with seedling phenotypes, which were all high infection types under normal temperatures. Variation in high infection types was observed among the susceptible lines ranging from Stakman infection type '3' to '3+'. Under lower temperature conditions, 15 of the lines had low infection types ranging from ';13' to '3+;'. The resistance genes conferring these low infection types at the cool temperature treatment are not known. The seedling phenotypes converted to a numerical scale were weakly correlated with the genetic values for adult plant resistance, with correlations of 0.1 and 0.19 for the normal and cool treatments respectively. Both correlations were significant, with p-values of 0.049 and 3×10^{-4} for the normal and cool treatments, respectively.

Genome-wide association analysis

Eight markers were associated with stem rust resistance (Table 4.1). *csSr2_ KASPar*, explained 27% of the variation in the genotypic values. Both *csSr2* and *csSr2_ KASPar* are tightly linked to *Sr2* located on chromosome 3BS (Mago et al., 2011). Two other markers associated with stem rust resistance are known to be located on chromosome 3BS based on the Synthetic x Opata genetic map (Poland et al., 2012a). The remaining four associated markers have unknown

Table 4.1: Markers significantly associated with adult plant stem rust resistance

Marker	MAF†	p-value	Effect	r^2	Chromosome
csSr2_KASPar	0.29	$3.38\times^{-10}$	0.54	0.27	3BS
csSr2	0.16	$1.21\times^{-8}$	0.65	0.17	3BS
GBS_13164	0.19	$1.62\times^{-6}$	0.6	0.15	-
GBS_11008	0.29	$7.09\times^{-6}$	0.49	0.08	3BS
GBS_1863	0.20	$1.01\times^{-5}$	0.51	0.17	-
GBS_7565	0.3	$1.19\times^{-5}$	0.48	0.07	-
GBS_10286	0.12	$2.83\times^{-5}$	-0.61	0.08	-
GBS_20803	0.32	$4.27\times^{-5}$	0.42	0.19	3BS

† Minor allele frequency

map location. Pairwise associations between significant markers, measured in r^2 indicated that two markers of unknown map location are associated, $r^2 \geq 0.4$, with markers known to be on chromosome 3B (Figure 4.2). The two remaining markers of unknown location are not associated with each other or other significant markers.

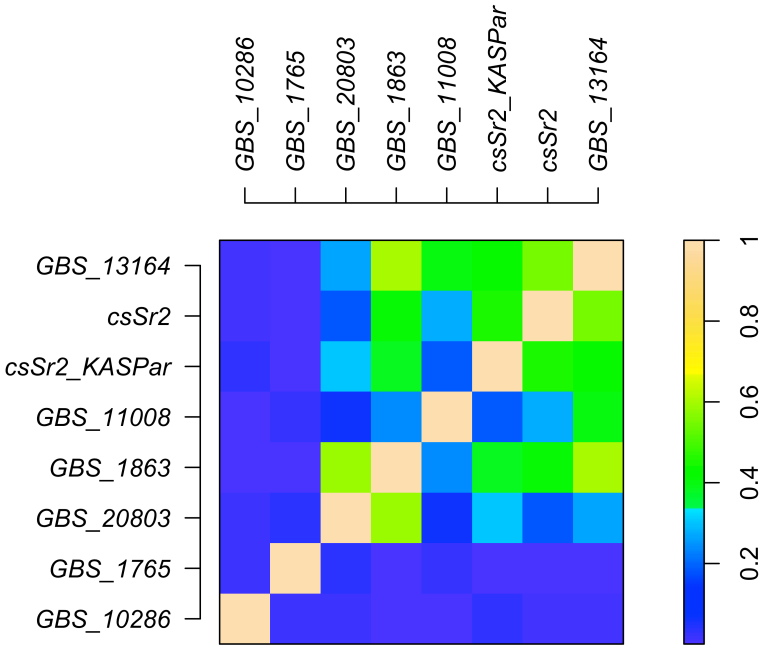


Figure 4.2: Pairwise associations, measured in r^2, between markers significantly

associated with adult plant stem rust resistance

The marker relationship matrix shows several small groups of closely related individuals, indicating family structure (Figure 4.3). Based on pedigree information, 147 of the individuals were derived from 26 full-sib families. Individuals derived from the same full sib-family were found to group together based on the relationship matrix (Figure 4.3). Principal components analysis of the relationship matrix also illustrated a similar pattern of family relationships (Figure 4.4); however principal component one and two explained only 14.4% and 2.9% of the variation respectively. Correcting for kinship during genome-wide association was necessary to obtain uniformly distributed p-values (Figure 4.5). Further correcting for population structure using principal components did not improve uniformity of p-values.

Prediction model accuracies

The marker set containing all markers, both GBS and gene targeted markers, always resulted in higher accuracies than the marker set containing only GBS markers based on accuracies calculated using cross validation (Table 4.2) and bootstrapping (Table 4.3). Among the GS models, G-BLUP B and G-BLUP A lead to the highest cross validation prediction accuracies, followed by G-BLUP T, BL, and BCπ. Based on a bootstrap significance testing procedure, probabilities that pairs of model accuracies were different due to chance (p-values) for all models except BL and BCπ were estimated (Table 4.3). For comparisons between G-BLUP B, and ordinary G-BLUP or MLR models, p-values were always less than

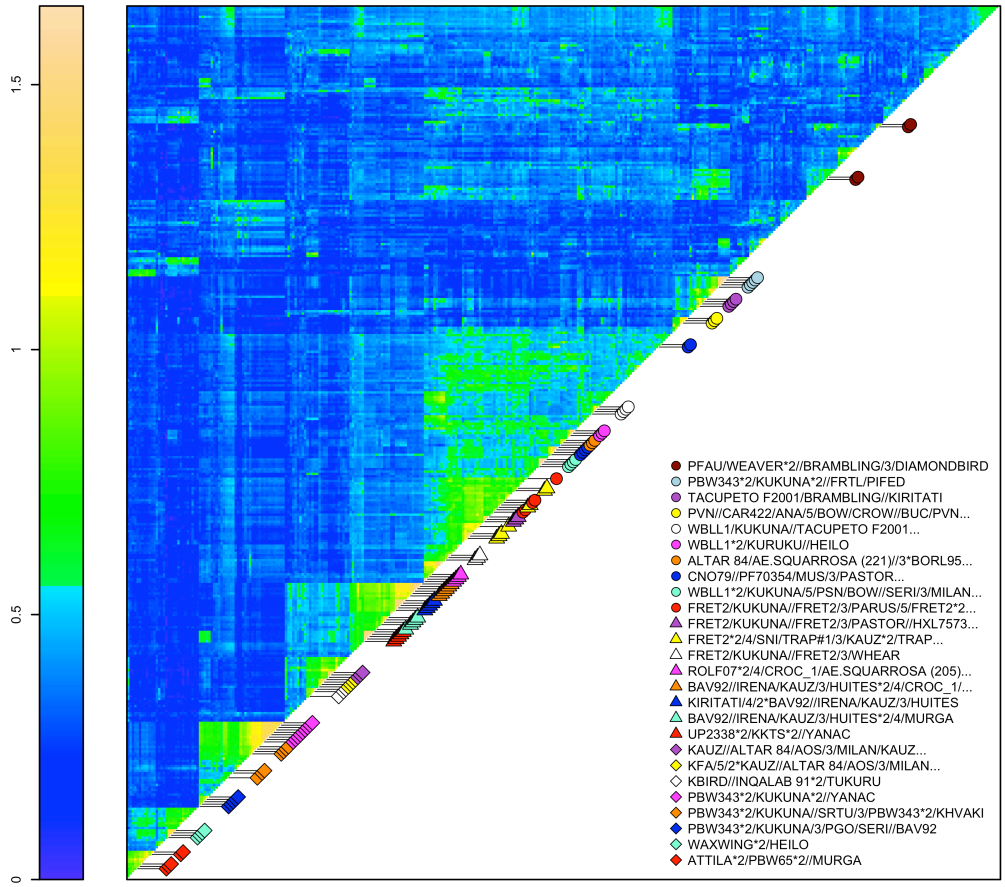


Figure 4.3: Heatmap of the marker relationship matrix illustrating family structure. Individuals derived from the same full-sib family share a common symbol.

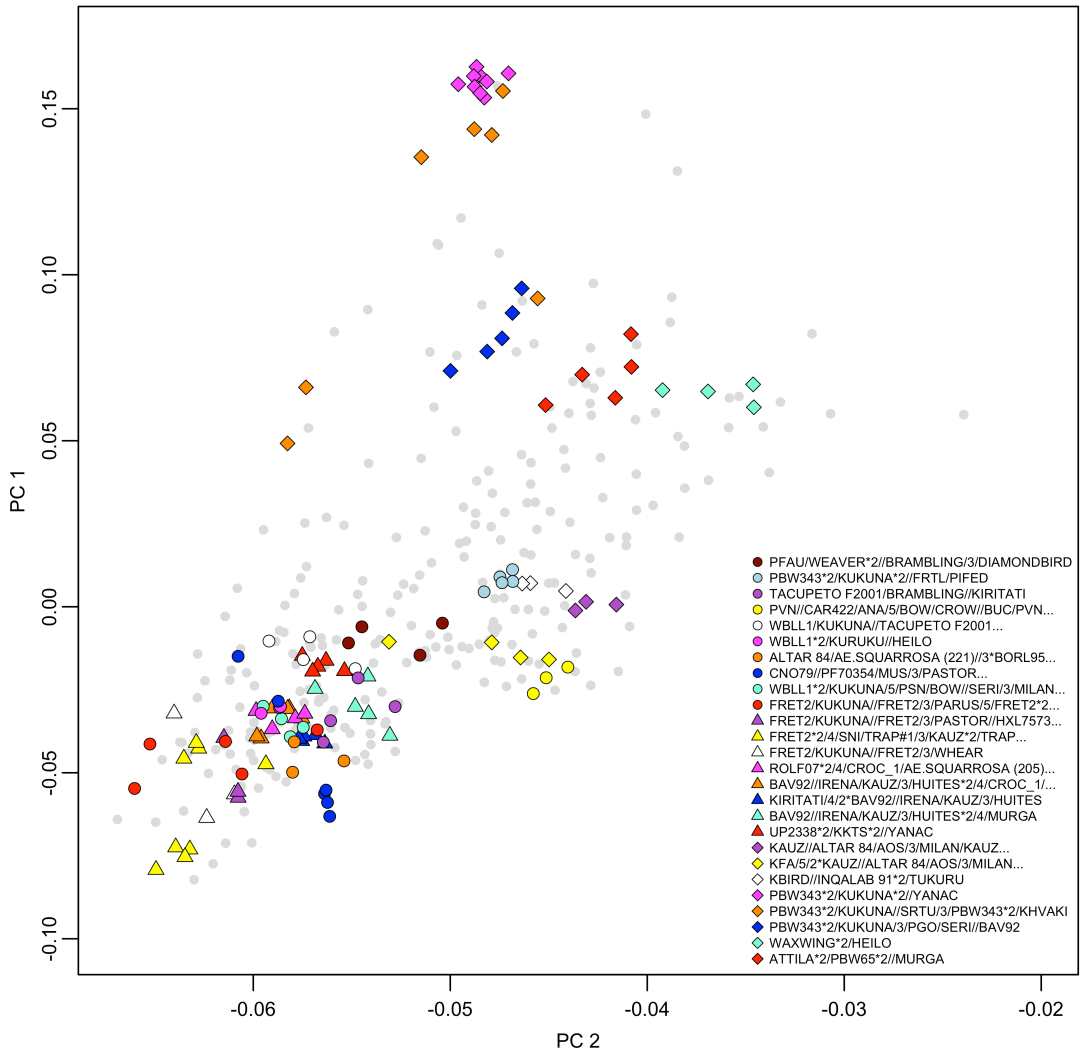


Figure 4.4: Principal components analysis of the marker relationship matrix. Individuals derived from the same full-sib family share a common symbol.

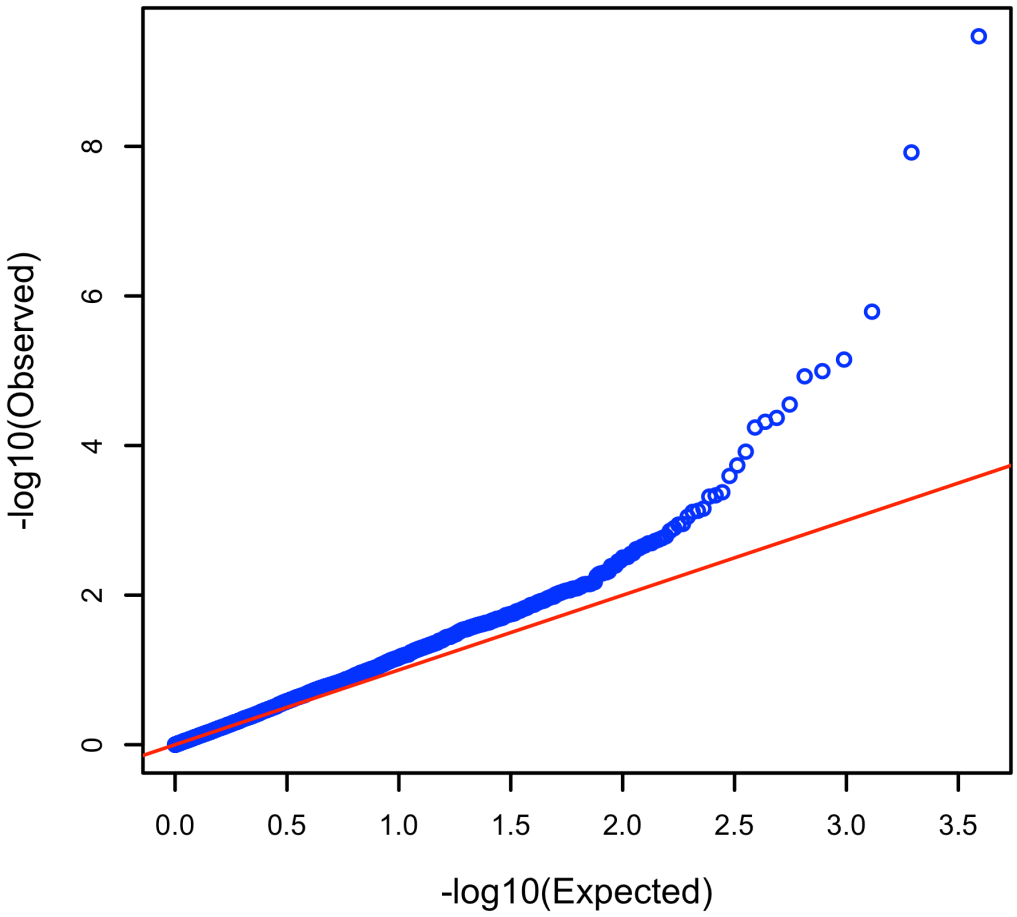


Figure 4.5: Quantile-quantile plot of the p-values from genome-wide association comparing the p-value distribution to a uniform null distribution

Table 4.2: Cross validation prediction accuracies for adult plant stem rust resistance using different prediction models and marker sets

Prediction Model†	All markers	GBS markers only
MLR A	0.477	0.446
MLR B	0.468	-
G-BLUP A	0.607	0.577
G-BLUP B	0.618	-
G-BLUP	0.568	0.563
BL	0.579	0.561
BCπ	0.578	0.558
G-BLUP T	0.591	0.573

† MLR A: Multiple linear regression A, fixed effects selected among all markers; MLR B, fixed effects selected among candidate gene linked markers; G-BLUP A: Genomic best linear unbiased prediction A, marker relationship matrix and fixed effects selected among all markers; G-BLUP B, marker relationship matrix and fixed effects selected among candidate gene linked markers; BL: Bayesian Lasso; BCπ: Bayes Cπ; G-BLUP T, marker relationship matrix and seedling phenotypes as fixed effects.

Table 4.3: Probabilities that pairs of model accuracies are not different based on bootstrapping.

		GBS Markers Only					All Markers				
	Model†, Accuracy	G-BLUP, 0.58	G-BLUP A, 0.54	G-BLUP T, 0.57	MLR A, 0.36	G-BLUP, 0.59	G-BLUP A, 0.63	G-BLUP B, 0.66	G-BLUP T, 0.58	MLR A, 0.51	MLR B, 0.56
GBS Markers Only	G-BLUP, 0.58	1	0.52	0.95	0.08	0.69	0.39	0.12	0.99	0.57	0.79
	G-BLUP A, 0.54	0.52	1	0.84	0.1	0.43	0.14	0.07	0.79	0.82	0.9
	G-BLUP T, 0.57	0.95	0.84	1	0.21	0.89	0.65	0.47	0.84	0.72	0.88
	MLR A, 0.36	0.08	0.1	0.21	1	0.08	0.02	0.01	0.18	0.15	0.08
All Markers	G-BLUP, 0.59	0.69	0.43	0.89	0.08	1	0.44	0.15	0.91	0.51	0.67
	G-BLUP A, 0.63	0.39	0.14	0.65	0.02	0.44	1	0.62	0.68	0.15	0.34
	G-BLUP B, 0.66	0.12	0.07	0.47	0.01	0.15	0.62	1	0.5	0.09	0.12
	G-BLUP T, 0.58	0.99	0.79	0.84	0.18	0.91	0.68	0.5	1	0.68	0.84
	MLR A, 0.51	0.57	0.82	0.72	0.15	0.51	0.15	0.09	0.68	1	0.75
	MLR B, 0.56	0.79	0.9	0.88	0.08	0.67	0.34	0.12	0.84	0.75	1

† G-BLUP A: Genomic best linear unbiased prediction A, marker relationship matrix and fixed effects selected among all markers; G-BLUP T: marker relationship matrix and seedling phenotypes as fixed effects; MLR A: Multiple linear regression A, fixed effects selected among all markers; G-BLUP B: marker relationship matrix and fixed effects selected among candidate gene linked markers; MLR B: Multiple linear regression B, fixed effects selected among candidate gene linked markers.

0.15. The markers that were selected in MLR A, and G-BLUP A were *csSr2_KASPar*, *GBS_20803*, *csSr2, GBS_1863* (Table 4.4).

Table 4.4: Markers used as fixed effects in different prediction models, their MAFs, and the frequency they appeared in the models during cross-validation

		Frequency Selected as Fixed Effects			
Marker	MAF†	MLR A‡	MLR B§	G-BLUP A¶	G-BLUP B#
*csSr2*_KASPar	0.29	1	1	1	1
csSr2	0.16	.5	1	.8	1
GBS_20803	0.31	.9	-	.6	-
csLV34	0.37	0	.4	0	0
gwm533	0.34	0	.2	0	0
GBS_1863	0.20	.1	0	0	0

† Minor allele frequency
‡ Multiple linear regression A, fixed effects selected among all markers
§ Multiple linear regression B, fixed effects selected among candidate gene linked markers
¶ Genomic best linear unbiased prediction A, marker relationship matrix and fixed effects selected among all markers
# Genomic best linear unbiased prediction B, marker relationship matrix and fixed effects selected among candidate gene linked markers

The map locations of *GBS_20803* and *GBS_1863* are unknown. The markers selected by G-BLUP B, the most accurate model, were *csSr2_KASPar* and *csSr2*. Differences in prediction model outcomes between pairs of prediction models are shown by their Spearman's rank correlations between estimated breeding values from cross validation for all pairs of models shown in Table 4.5. MLR B had the lowest correlations between all other models followed by MLR A.

Discussion

Genetic architecture

The association analysis results confirm the importance of the *Sr2* region, with the most significant *Sr2* linked marker explaining 27% of the variation. Out of

Table 4.5: Spearman's rank correlations between estimated breeding values for all pairs of model

		GBS Markers Only							All Markers						
	Model	BCπ	BL	G-BLUP A	G-BLUP	G-BLUP T	MLR A	BCπ	BL	G-BLUP	G-BLUP A	G-BLUP B	G-BLUP T	MLR A	MLR B
GBS Markers Only	BCπ	1	1	0.91	0.96	0.95	0.63	0.97	0.99	0.96	0.88	0.88	0.95	0.52	0.41
	BL	1	1	0.91	0.96	0.96	0.63	0.97	0.99	0.97	0.88	0.88	0.95	0.52	0.41
	G-BLUP A	0.91	0.91	1	0.93	0.89	0.78	0.91	0.91	0.92	0.9	0.89	0.89	0.61	0.46
	G-BLUP	0.96	0.96	0.93	1	0.95	0.65	0.93	0.95	0.98	0.89	0.89	0.94	0.54	0.42
	G-BLUP T	0.95	0.96	0.89	0.95	1	0.63	0.93	0.95	0.95	0.86	0.87	0.98	0.52	0.42
	MLR A	0.63	0.63	0.78	0.65	0.63	1	0.65	0.64	0.65	0.68	0.66	0.65	0.65	0.48
All Markers	BCπ	0.97	0.97	0.91	0.93	0.93	0.65	1	0.98	0.94	0.93	0.94	0.94	0.59	0.51
	BL	0.99	0.99	0.91	0.95	0.95	0.64	0.98	1	0.96	0.91	0.91	0.95	0.56	0.46
	G-BLUP	0.96	0.97	0.92	0.98	0.95	0.65	0.94	0.96	1	0.9	0.91	0.95	0.55	0.44
	G-BLUP A	0.88	0.88	0.9	0.89	0.86	0.68	0.93	0.91	0.9	1	0.99	0.88	0.7	0.61
	G-BLUP B	0.88	0.88	0.89	0.89	0.87	0.66	0.94	0.91	0.91	0.99	1	0.89	0.68	0.61
	G-BLUP T	0.95	0.95	0.89	0.94	0.98	0.65	0.94	0.95	0.95	0.88	0.89	1	0.55	0.44
	MLR A	0.52	0.52	0.61	0.54	0.52	0.65	0.59	0.56	0.55	0.7	0.68	0.55	1	0.72
	MLR B	0.41	0.41	0.46	0.42	0.42	0.48	0.51	0.46	0.44	0.61	0.61	0.44	0.72	1

† BCπ: Bayes Cπ; BL: Bayesian Lasso; G-BLUP A: Genomic best linear unbiased prediction A, marker relationship matrix and fixed effects selected among all markers; G-BLUP T, marker relationship matrix and seedling phenotypes as fixed effects; MLR A: Multiple linear regression A, fixed effects selected among all markers; G-BLUP B, marker relationship matrix and fixed effects selected among candidate gene linked markers; MLR B, fixed effects selected among candidate gene linked markers

eight significant markers, only two markers did not appear to be at the *Sr2* region. *Sr2* linked markers have been reported by several stem rust adult plant resistance studies (Yu et al., 2011; Njau et al., 2012; Singh et al., 2013). Interestingly, the most significant *Sr2* linked marker was *csSr2_ KASPar*. This marker gave different results than the STS marker of *csSr2*, which has been reported to not be diagnostic for *Sr2* in CIMMYT germplasm (Mago et al., 2011). Our results suggest that *csSr2_ KASPar* is capturing a different haplotype than the *csSr2* STS marker. This may be due to restriction site polymorphism at the restriction enzyme cut site of the STS marker. Marker *gwm533*, which is still used for *Sr2* genotyping, was not associated with resistance in this study, suggesting that this marker should be discontinued for *Sr2* genotyping. In contrast with other studies (Dyck, 1987; Krattinger et al., 2009; Singh et al., 2012), this study did not find *Lr34* to be associated with adult plant stem rust resistance . The frequency of the *Lr34* resistance allele was 0.36, thus the lack of association between *Lr34* and resistance was not due to low minor allele frequency. In the association mapping study by Yu et al. (2011), which used a similar set of germplasm and environments, *Lr34* was also not found to be significant; however several significant marker interactions with *Lr34* were detected. Based on the inconsistencies in detection and the reported marker interactions, the effect of *Lr34* appears to vary depending on the genetic background.

The relatively low number of QTLs that we detected is due largely to the confounding of QTL effects with family structure. Without correcting for

population or family structure, 138 markers exceed the significance threshold and the p-values do not follow a uniform distribution, indicating many spurious associations. Confounding of marker effects with family structure is not a problem for GS because GS capitalizes on relationship information to predict breeding values.

Prediction models

A G-BLUP model including *Sr2* linked markers as fixed effects was the most accurate model tested, and the probability that this model was different from MLR with *Sr2* linked markers alone and G-BLUP with GBS markers only was 0.12 and 0.15 respectively. These results suggest that GS based on G-BLUP with *Sr2* linked markers as fixed effects would lead to the greatest genetic gain if GS was imposed on the specific dataset used in this study. However, if GS were to be applied on a new sample of individuals, there is some probability that the outcomes of GS using G-BLUP with GBS markers only, or MLR using *Sr2* linked markers only would be just as favorable as the outcomes of GS using G-BLUP with *Sr2* linked markers as fixed effects.

Our finding that modeling selected markers as fixed effects in G-BLUP leads to improved accuracy over standard G-BLUP agrees with a recent simulation study (Bernardo, 2013) which found modeling a large-effect locus as fixed to be advantageous when heritability of the trait was greater than 0.5 and the proportion of the genetic variance explained by the locus was greater than 0.25. It is important to emphasize that in this study, the markers selected as fixed

effects were not assumed to be causative loci, thus variable selection and fixed effect estimation should occur each time the prediction model is trained.

The correlation between low temperature seedling and adult plant phenotypes was interesting, but not sufficient to be useful in combination with GS in the germplasm tested. Using the seedling data as fixed effects in G-BLUP did not consistently improve the prediction accuracy. Seedling data could be more predictive in another set of germplasm. On the other hand, if the level of adult plant resistance can be explained well by seedling infection types, the resistance may be mostly qualitative, due to single race-specific genes. Thus, it may not be desirable to use this information source even if it is predictive of adult-stage resistance.

If we assume that two cycles of GS can be completed for every one cycle of phenotypic selection, and all other factors remain constant, then gain from selection from GS will exceed the gain from phenotypic selection when (GS accuracy $\times$ 2) is greater than the phenotypic selection accuracy. The GS accuracies we achieved in this study are sufficiently high to achieve greater gain from selection per unit time compared to phenotypic selection. Phenotypic selection accuracy, estimated as $\sqrt{H^2}$, was 0.9, and (GS accuracy $\times$ 2) was 1.12. The GS accuracies we observed were similar to those observed in a GS study that evaluated prediction accuracies for stem rust resistance in bi-parental populations (Ornella et al., 2012), however the results are difficult to compare due to different training population sizes.

Conclusion

This study indicates that GS would be an effective breeding method for quantitative stem rust resistance despite the fact that the trait is highly heritable and is conferred in part by large-effect loci. Although one of the advantages of GS is that prior knowledge about loci affecting the trait is not needed, we found that in this dataset using prior information to selectively genotype markers at loci previously found to have a moderately large effect on the trait enabled us to achieve higher prediction accuracies especially when using models which treat large-effect loci as fixed effects. To ensure the best results from GS, markers linked to large to moderate effect genes or loci previously found to affect the traits of interest should be included in the genotypic data as long as doing so does not delay selection or incur excessive costs. Using cross-validation within the training data, one can then decide if these loci specific markers should be modeled as fixed effects. Although the alleles at 'known' loci may be different from those of the population where the loci were detected, they may still be important regions that should be tagged with markers. As more genes are mapped and cloned in wheat for various traits, the effect of utilizing gene information for genomic prediction of other traits in wheat can be further studied.

Acknowledgements

This research was funded by The Bill & Melinda Gates Foundation (Durable Rust Resistance in Wheat) and the United States Department of Agriculture[1]-

Agricultural Research Service (USDA-ARS) (Appropriation No. 5430-21000-006-00D). Partial support for J. Rutkoski was provided by a USDA National Needs Fellowship Grant #2008- 38420-04755 and an American Society of Plant Biology (ASPB) -Pioneer Hi-Bred Graduate Student Fellowship. Loci targeted genotyping was provided by the Eastern Regional Small Grains Genotyping Laboratory, Raleigh, North Carolina. Statistical advice was provided by the Cornell Statistical Consulting Unit.

References

Bates, D., and M. Maechler. 2010. lme4: Linear mixed-effects models using S4 classes. Available at http://cran.r-project.org/package=lme4 (accessed 7 Feb. 2014).

Bernardo, R. 1994. Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci. 34: 20–25.

Bernardo, R. 2013. Genomewide selection when major genes are known. Crop Sci. 2014. 54:68-75.

Box, G.E.., and D.R. Cox. 1964. An analysis of transformations. J. R. Stat. Soc. Ser. B 26: 211–252.

Carlson, C.S., M. a Eberle, M.J. Rieder, Q. Yi, L. Kruglyak, and D. a Nickerson. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am. J. Hum. Genet. 74: 106–120.

Dyck, P.L. 1987. The association of a gene for leaf rust resistance with the chromosome 7D suppressor of stem rust resistance in common wheat. Genome 29: 1986–1988.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species (L Orban, Ed.). PLoS One 6: e19379.

Endelman, J. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Gen. 4: 250-255.

Fraley, C., and A.E. Raftery. 2002. Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. 97 611–631.

Fraley, C., A.E. Raftery, T.B. Murphy, and L. Scrucca. 2012. Mclust version 4 for r: normal mixture modeling for model-based clustering, classification, and density estimation. Avaliable at http://www.stat.washington.edu/research/reports/2012/tr597.pdf (accessed 7 Feb. 2014).

Garrick, D.J., J.F. Taylor, and R.L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41: 55.

Habier, D., R.L. Fernando, K. Kizilkaya, and D.J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12: 186.

Hallauer, A.R., M.J. Carena, and J.B. Miranda Filho. 2010. Quantitative genetics in maize breeding. Iowa State University Press, Ames, IA.

Heffner, E.L., M.E. Sorrells, and J.-L. Jannink. 2009. Genomic selection for crop improvement. Crop Sci. 49: 1-12.

Jin, Y., U. States, A. Agricultural, U.C.D. Labo-, R. Wanyera, M. Kinyua, P. Njau, and K. Agri-. 2007. Characterization of seedling infection types and adult plant infection responses of monogenic Sr gene lines to race TTKS of *Puccinia graminis* f . sp . *tritici.* 91: 1096–1099.

Kang, H.M., J.H. Sul, S.K. Service, N.A. Zaitlen, S. Kong, N.B. Freimer, C. Sabatti, and E. Eskin. 2010. Variance component model to account for sample structure in genome-wide association studies. Nat. Publ. Gr. 42: 348–354.

Krattinger, S.G., E.S. Lagudah, W. Spielmeyer, R.P. Singh, J. Huerta-Espino, H. McFadden, E. Bossolini, L.L. Selter, and B. Keller. 2009. A putative ABC transporter confers durable resistance to multiple fungal pathogens in wheat. Sci. 323: 1360–1363.

Lagudah, E.S., S.G. Krattinger, S. Herrera-Foessel, R.P. Singh, J. Huerta-Espino, W. Spielmeyer, G. Brown-Guedira, L.L. Selter, and B. Keller. 2009. Gene-specific markers for the wheat gene Lr34Yr18/Pm38/ which confers resistance to multiple fungal pathogens. Theor. Appl. Genet. 119: 889–898.

Lagudah, E.S., H. McFadden, R.P. Singh, J. Huerta-Espino, H.S. Bariana, and W. Spielmeyer. 2006. Molecular genetic characterization of the Lr34/Yr18 slow rusting resistance gene region in wheat. Theor. Appl. Genet. 114: 21–30.

Lipka, A.E., F. Tian, Q. Wang, J. Peiffer, M. Li, P.J. Bradbury, M.A. Gore, E.S. Buckler, and Z. Zhang. 2012. GAPIT: genome association and prediction integrated tool. Bioinformatics 28: 2397–2399.

Lorenz, A.J., S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata, K.P. Smith, M.E. Sorrells, and J. Jannink. 2011. Genomic selection in plant breeding : Knowledge and prospects. Adv. Agron. 110: 77–123.

De los Campos, G., and P. Perez Rodriguez. 2010. BLR: Bayesian linear regression. Available at http://cran.r-project.org/package=BLR (accessed 7 Feb. 2014).

Mago, R., G. Brown-Guedira, S. Dreisigacker, J. Breen, Y. Jin, R. Singh, R. Appels, E.S. Lagudah, J. Ellis, and W. Spielmeyer. 2011. An accurate DNA marker assay for stem rust resistance gene Sr2 in wheat. Theor. Appl. Genet. 122: 735–744.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Njau, P.N., S. Bhavani, J. Huerta-Espino, B. Keller, and R.P. Singh. 2012. Identification of QTL associated with durable adult plant resistance to stem rust race Ug99 in wheat cultivar “Pavon 76.” Euphytica 190: 33–44.

Ornella, L., S. Singh, P. Perez, J. Burgueño, R. Singh, E. Tapia, S. Bhavani, S. Dreisigacker, H.-J. Braun, K. Mathews, and J. Crossa. 2012. Genomic prediction of genetic values for resistance to wheat rusts. Plant Gen. 5: 136–148.

Park, R.F. 2007. Stem rust of wheat in Australia. Aust. J. Agric. Res. 58: 558–566.

Park, T., and G. Casella. 2008. The Bayesian Lasso. J. Am. Stat. Assoc. 103: 681–686.

Parlevliet, J.E. 2002. Durability of resistance against fungal, bacterial and viral pathogens; present situation. Euphytica 124: 147–156.

Pérez, P., G.D.L. Campos, J. Crossa, D. Gianola, and G. de los Campos. 2010. Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. Plant Gen. 3: 106–116.

Peterson, R.F., A.B. Campbell, and A.E. Hannah. 1948. A diagrammatic scale for estimating rust intensity on leaves and stems of cereals. Can. J. Res. 26c: 496–500.

Piepho, H.P. 2009. Ridge regression and extensions for genomewide selection in maize. Crop Sci. 49: 1165–1176.

Poland, J. A., P.J. Brown, M.E. Sorrells, and J.-L. Jannink. 2012a. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7: e32253.

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sánchez-Villeda, M. Sorrells, and J.-L. Jannink. 2012b. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Gen. 5: 103–113.

Pretorius, Z.A., R.P. Singh, W.W. Wagoire, and T.S. Payne. 2000. Detection of virulence to wheat stem rust resistance gene Sr31 in *Puccinia graminis*. f. sp. *tritici* in Uganda. Plant Dis. 84: 203.

R Development Core Team. 2010. R: A Language and environment for statistical computing. Available at http://www.r-project.org/ (accessed 7 Feb. 2014).

Rutkoski, J.E., J. Poland, J.-L. Jannink, and M.E. Sorrells. 2013. Imputation of unordered markers and the impact on genomic selection accuracy. G3 (Bethesda). 3: 427–439.

Singh, R.P., S. Herrera-Foessel, J. Huerta-Espino, H. Bariana, U. Bansal, B. Mccallum, C. Hiebert, S. Bhavani, S. Singh, C. Lan, and E. Lagudah. 2012. Lr34/Yr18/Sr57/ Pm38/Bdv1/Ltn1 Confers slow rusting, adult plant resistance to *Puccinia graminis tritici*. *In* Chen, W.-Q. (ed.), Proceedings of the 13th International Cereal Rusts and Powdery Mildews Conference. Beijing, China.

Singh, R.P., D.P. Hodson, Y. Jin, J. Huerta-Espino, M.G. Kinyua, R. Wanyera, P. Njau, and R.W. Ward. 2006. Current status, likely migration and strategies to mitigate the threat to wheat production from race Ug99 (TTKS) of stem rust pathogen. 1: 1–13.

Singh, S., R.P. Singh, S. Bhavani, J. Huerta-Espino, and E.E. Lopez-Vera. 2013. QTL mapping of slow-rusting, adult plant resistance to race Ug99 of stem rust fungus in PBW343/Muu RIL population. Theor. Appl. Genet. 126: 1367–1375.

Spielmeyer, W., P.J. Sharp, and E.S. Lagudah. 2003. Identification and validation of markers linked to broad-apectrum stem rust resistance gene Sr2 in wheat (*Triticum aestivum* L.). CRC. Crit. Rev. Plant Sci. 43: 333–336.

Stakman, E.C., D.M. Steward, and W.Q. Loegering. 1962. Identification of physiologic races of *Puccinia graminis* var. *tritici*. U.S. Dep. Agric. Res. Serv. E-617.

Sunderwirth, S.D., and A.P. Roelfs. 1980. Greenhouse evaluation of the adult-plant resistance of Sr2 to wheat-stem rust. Phytopathology 70: 634–637.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423.

Yu, L.-X., A. Lorenz, J. Rutkoski, R.P. Singh, S. Bhavani, J. Huerta-Espino, and M.E. Sorrells. 2011. Association mapping and gene-gene interaction for stem rust resistance in CIMMYT spring wheat germplasm. Theor. Appl. Genet. 123: 1257–1268.

Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38: 203–208.

Zhang, D., R. Bowden, and G. Bai. 2011. A method to linearize Stakman infection type ratings for statistical analysis. p. 170. *In* Borlaug Global Rust Initiative 2011 Technical Workshop June 13-16 Saint Paul, Minnesota, USA.

CHAPTER 5

EFFICIENT USE OF HISTORICAL DATA FOR GENOMIC SELECTION: A CASE STUDY IN WHEAT[5]

Abstract:

Genomic selection (GS) is a new methodology that can improve wheat breeding efficiency. To implement GS, a training population (TP) with both phenotypic and genotypic data is required to train a statistical model used to predict genotyped selection candidates (SCs). Several factors impact prediction accuracy, the relationship between the TP and the SCs being one of the most important. This study investigated the utility of a historical TP_H compared with a population specific TP, the potential for TP optimization using historical TP_H subsets, and the utility of historical TP data when close relative data is available for training. We found that, depending on TP size, a population specific TP was 1.5 to 4.4 times more accurate than a historical TP. TP optimization based on the mean of the generalized coefficient of determination (CDmean) or prediction error variance (PEVmean) enabled the selection of historical TP subsets that were significantly more accurate than randomly selected subsets. Retaining historical data when data on close relatives were available lead to a 11.9%

[5] A paper to be submitted to The Plant Genome as: Rutkoski J. E., J. Poland, R.P. Singh, J. Huerta-Espino, S. Bhavani, J-L. Jannink, M. E. Sorrells. Efficient use of historical data for genomic selection: A case study in wheat. Plant Gen.

increase in accuracy at best, and a 12% decrease in accuracy at worst depending on the heritability. We conclude that historical data could be used successfully to initiate a GS program, especially if the dataset is very large and of high heritability. TP optimization would be useful for the identification of historical TP subsets to phenotype additional traits. However after model updating, discarding historical data may be warranted. More empirical studies are needed to determine if these observations represent general trends.

Abbreviations

APR, Adult plant resistance; BLUP, Best linear unbiased prediction; CDmean, Mean of the generalized coefficient of determination; FA, factor analytic; G-BLUP, Genomic best linear unbiased prediction; GBS, Genotyping-by-sequencing; GS, Genomic selection; GxE, Genotype by environment interaction; LD, Linkage disequilibrium; MAF, Minor allele frequency; PC, Principal component; PEVmean, Mean of the prediction error variance; QTL, Quantitative trait loci; REML, Restricted estimation maximum likelihood, SCs, Selection candidates; TP, Training population; $TP_{H,}$ Historical training population; $TP_{PS,}$ Population specific training population.

Introduction

Genomic selection (GS) (Haley and Visscher, 1998; Meuwissen et al., 2001) is a breeding methodology that can increase rates of genetic gain by reducing the breeding cycle duration or by increasing the selection accuracy. With GS, a training population (TP) consisting of phenotyped and genotyped individuals is

used to train a model that predicts breeding values of genotyped selection candidates (SCs). The accuracy of this prediction depends on the TP size (*Np*), heritability (h^2), effective number of loci (*Me*), and the level of linkage disequilibrium (LD) between genetic markers and quantitative trait loci (QTL) (Goddard, 2009; Daetwyler et al., 2010). If the TP and SCs are from different populations, the genetic relationship between these two populations is another major factor affecting GS accuracy (Habier et al., 2007; de Roos et al., 2009; Hayes et al., 2009; Long et al., 2011; Pszczola et al., 2012). As the relationship between the TP and SC decreases, the forces of selection, recombination, and drift, change the pattern of LD between markers and QTL. Furthermore, markers that capture family effects rather than QTL effects contribute much less to the GS accuracy as relationship between the TP and SCs declines (Habier et al., 2007). The non-additivity of QTL effects may also contribute to the decrease in accuracy as the relationship between the TP and SCs decreases.

In plant breeding, there is considerable interest in the use of historical data for GS model training to predict breeding values of new SCs (Crossa et al., 2010; Asoro et al., 2011; Storlie and Charmet, 2013; Dawson et al., 2013). Compared to a 'population specific' TP that consists of a subset of the SCs selected for model training, a historical TP enables predictions to be generated sooner in the breeding cycle because phenotyping the TP occurs before the selection candidates are developed. In addition, a historical TP could be of higher line mean heritability and sample more environments compared to a newly

generated population specific TP. On the other hand, compared to a population specific TP, a historical TP_H consists of more distant relatives, which can lead to reduced accuracy. Studies which have assessed GS accuracy from historical data in crop species have measured accuracy using either random cross-validation or 'forward validation', where an older subset of the data is used to predict a newer subset (Asoro et al., 2011; Dawson et al., 2013). Accuracies from random cross-validation are likely to be over estimated because the TP and SCs are from the same population. On the other hand, accuracies from forward validation may be driven largely by the level of genotype by environment interaction (GxE) between the training and validation environments rather than the relationship between the TP and SCs. As a result, there are no empirical studies that can clearly demonstrate the utility of historical data for the prediction of new SCs assuming that the historical set of environments represent those environments of interest to the breeding program.

The purpose of this case study was to assess the utility of historical data for the prediction of new, early generation SCs. We used empirical data from a recurrent genomic selection program for stem rust (*Puccinia graminis* f. sp. *tritici*) adult plant resistance in wheat (*Triticum aestivum*) to 1) determine the relative accuracies of historical and 'population specific' training sets for the prediction of new SCs, 2) determine the potential to use TP optimization methods to identify the best subsets of historical individuals to use for training and 3) determine if historical data should remain part of the TP if data on close relatives

becomes available for model training.

Materials and Methods

Genetic material

A set of three hundred sixty five advanced CIMMYT wheat lines was used as the historical population. A second population of five hundred three new SCs was generated by two rounds of random mating between fourteen founder lines from the historical population, followed by one round of selfing for seed increase. Each SC was phenotypically evaluated based on its S_1 or S_2 progeny. Each SC was genotyped using bulk DNA from six progenies.

Phenotypic data

Individuals were evaluated for quantitative adult plant resistance (APR) to stem rust at the Kenya Agricultural Research Institute, Njoro, Kenya and/or the Ethiopian Institute of Agricultural Research, Debre Zeit, Ethiopia as described in Yu et al. (2011). The historical population was evaluated across 10 seasons in Kenya and three seasons in Ethiopia from 2007 and 2013, with each individual appearing in approximately four of the 13 environments. The SCs were evaluated in Kenya during the 2012 main and off-season and during the 2013 main-season where they were planted in twin row field plots of 70cm and 30cm spacing surrounded by a 1m border of spreader plants. Hills of spreader plants were planted in in rows perpendicular to the entry rows. Just prior to booting (growth stage Z35- Z37; Zadoks et al. 1974) individual spreader plants of the border rows were inoculated with fresh urediniospores of *Puccinia graminis* f. sp. *tritici* race

TTKST (Sr24 virulent race) suspended in distilled water using a hypodermic syringe, on at least two occasions. Spreaders were also sprayed with suspension of urediniospores in light mineral oil Soltrol 170 to ensure successful infection. Disease severity was determined according to modified Cobb scale (Peterson et al., 1948), and a Box-Cox transformation (Box and Cox, 1964) was applied prior to analysis. For both the historical and selection candidate populations heritability on a line mean basis was calculated according to Hallauer et al. (2010). Variance components were estimated using the R package *lme4* (Bates and Maechler, 2010).

Genotypic data

Genotyping-by-sequencing (GBS, Elshire et al., 2011) was implemented according to the protocol described in (Poland et al., 2012a). Out of the total of 27,434 polymorphic markers generated, 17,168 unique markers with less than 80% missing data in the historical population, and polymorphic in the selection candidates were selected. Prior to marker filtering, missing data was imputed using random forest imputation described in Poland et al. (2012b) as recommended by Rutkoski et al. (2013).

Relationship matrix

The relationship matrix (**A**) was calculated according to Leutenegger et al. (2003), Amin et al. (2007), and Astle and Balding (2009). Relationship estimates for a pair of individuals *i* and *j* was:

$$f_{ij} = \frac{1}{n}\sum_{k=1}^{n}\frac{(g_{ik} - p_k)(g_{jk} - p_k)}{p_k(1 - p_k)}$$

g_{ik} is the genotype of individual *i* at marker *k* coded as 0, 0.5, 1, p_k is the frequency of the major allele, and *n* is the number of markers used for kinship estimation. Prior to relationship matrix calculation, markers with a minor allele frequency (MAF) less than 0.05 were excluded.

Population characterization

Population differentiation between the three hundred sixty five historical lines and the five hundred three SCs was measured using the Fst (Wright, 1949):

$$\text{Fst} = \frac{H_T - \bar{H}_S}{H_T}$$

$H_T = 2\bar{p}\bar{q}$, where $\bar{p}$ and $\bar{q}$ are the weighted average of the within sub-population allele frequencies *p* and *q*, and $\bar{H}_S$ is the weighted average of 2*pq* within sub-populations. For each marker only non-imputed data points were used. Statistical significance of the median Fst across all markers was assessed using 1000 permutations. For each iteration, the population assignment of the individuals was randomly shuffled prior to calculating median Fst. The distribution of the 1000 median Fst values was used as the null distribution for p-value calculation.

To visualize the population structure of the combined historical and selection candidate population, principal component (PC) analysis of the relationship matrix was implemented in R (R Development Core Team, 2010).

LD decay in historical and the SC population was investigated by plotting

the r^2 *vs.* genetic distance in centimorgans (cM) for pairs of markers on the same chromosome. Estimates of marker position from the Synthetic W9784 x Opata85 genetic map (Poland et al., 2012a) were available for 2425 markers. Markers with unknown map position and markers with MAF less than 0.05 were excluded, leaving 2050 markers. For each pairwise r^2 calculation, only non-imputed data points were used and marker pairs were excluded if there were less than thirty pairwise complete observations.

GS model

A single stage genomic best linear unbiased prediction (G-BLUP) model (Bernardo, 1994; Piepho, 2009), was used for all genomic predictions:

$$\mathbf{u} \sim \mathrm{N}(0, \mathbf{A}\sigma_{\mathrm{u}}^2)$$

$$\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \mathbf{R}\sigma_{\varepsilon}^2)$$

Y is a vector of phenotypes, **β** is a vector of environment effects treated as fixed, **u** is a vector of genotype effects treated as random, **X** and **Z** are the design matrices relating **β** and **u** to the observations in **Y**, **ε** is the residual error, σ_{u}^2 is the genetic variance, $\sigma_{\boldsymbol{\varepsilon}}^2$ is the error variance and **R** was the residual covariance matrix. **R** was equal to the identity matrix unless specified otherwise. The G-BLUP solutions for the breeding values were obtained using the mixed model equations (Henderson, 1984):

$$\begin{bmatrix} \mathbf{X}\mathbf{R}^{-1}\mathbf{X}' & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}\mathbf{R}^{-1}\mathbf{X} + \lambda\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix}$$

$\hat{\boldsymbol{\beta}}$ was the vector of fixed effect solutions, $\hat{\mathbf{u}}$ was the vector of estimated breeding values, and $\lambda = \frac{\hat{\sigma}^2_{\boldsymbol{\varepsilon}}}{\hat{\sigma}^2_{\mathbf{u}}}$. The variance components $\hat{\sigma}^2_{\boldsymbol{\varepsilon}}$ and $\hat{\sigma}^2_{\mathrm{u}}$ were estimated with the training set using restricted estimation maximum likelihood (REML) implemented in the R package *rrBLUP* (Endelman, 2011).

TP accuracy comparison

Out of the five hundred three SCs, one hundred thirty eight selected to be representative of the entire population based on pedigree were set aside as the validation population. The remaining three hundred sixty five SCs were set aside as the population specific TP (TP_{PS}). The 365 historical lines formed the historical TP (TP_H). TP_{PS} and TP_H were compared in terms of accuracy for *Np* values: 73, 146, 219, 292, and 365. For each accuracy calculation, 1000 random samples of size *Np* were used for model training, validation, and accuracy calculation. For each level of *Np*, a 95% confidence interval for accuracy was constructed by sorting the 1000 accuracies from smallest to largest and using the 24^{th} and 974^{th} accuracy values as the lower and upper confidence limits. Lastly, an average λ across the 1000 samples for each *Np* was computed (λ_{Np}) for use in later analyses.

The validation set was evaluated across two environments: Kenya main-season 2012 and Kenya main-season 2013. For model training, data from Kenya main-season 2012 and Kenya main-season 2013 were excluded so that the training and validation environments would not overlap. Accuracies are reported

as the Pearson's correlation between the G-BLUPs and the de-regressed genetic values of the validation set. To estimate the genetic values of the validation set, the R package *rrBLUP* (Endelman, 2011) was used to fit the mixed model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

$$\mathbf{u} \sim \mathrm{N}(0, \mathbf{I}\sigma_{\mathrm{u}}^2)$$

$$\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \mathbf{I}\sigma_{\varepsilon}^2)$$

where *I* is an identity matrix. Genetic values, u, were used for GS model validation .

Correlation between model training and validation environments

A factor analytic (FA) model, implemented in ASreml-R (Gilmour et al., 2009), was fit to parsimoniously model the covariance among environments. The FA model estimates the unobserved common factors, *k*, that give rise to the correlations between the environments, *e*. The environmental covariance matrix is modeled as:

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}$$

where $\boldsymbol{\Gamma}$ is an *e x k* matrix of factor loadings and $\boldsymbol{\Psi}$ is an *e x e* diagonal matrix of environment specific variances. FA variance models including genomic relationship information were fit for *k*=1, 2, and 3 according to (Beeck et al., 2010). Data from 16 environments between 2005 and 2012, was used to fit the FA models. The FA *k*=2 model was selected based on the Akaike information criterion (AIC). Estimates of variance parameters were used in the mixed model equations to estimate empirical BLUPs of each individual *i* in each environment *j*,

$\hat{u}_{ij}$ according to Thompson et al., (2003), which allows for variance matrices which are not full rank. The genetic value of each validation individual, *i* across a set of *N* environments was predicted as

$$\bar{u}_i = \frac{1}{N}\sum_{j}^{N}\hat{u}_{ij}$$

This was calculated for the validation individuals across the set of historical training environments, $\bar{\mathbf{u}}_H$, and across the set of population specific training environments $\bar{\mathbf{u}}_{PS}$. Correlations between each of the training environments and the set of validation environments were calculated as $cor(\mathbf{u}', \bar{\mathbf{u}}_H)$, and $cor(\mathbf{u}', \bar{\mathbf{u}}_{PS})$.

TP optimization

Two approaches were tested for TP optimization, 1) minimize the genetic differentiation between the training and validation populations or 2) maximize the precision of the predicted difference between each validation set individual and the mean of the validation population. For the first approach, the median Fst across all markers was the TP optimization criterion. For the second approach the mean PEV (PEVmean) and the mean coefficient of determination (CDmean) were tested as TP optimization criteria as suggested by Rincent et al. (2012). PEVmean and CDmean were recommended by Kennedy and Trus (1993) and Laloë (1993), respectively, as measures of the predictability of contrasts for breeding value estimation by best linear unbiased prediction (BLUP). Precise estimation of the contrasts (differences) between the overall selection candidate population mean and the individual breeding values is key for the identification

of the best individuals for selection.

For each population consisting of a potential training set of size *Np* and the validation set, according to Rincent et al. (2012), PEVmean was computed as

$$\sum_{i=1}^{N_v} \frac{c'_i(\mathbf{Z'MZ}+\lambda\mathbf{A^{-1}})^{-1}c'_i}{c'_i c_i} \times \hat{\sigma}^2_{\boldsymbol{\varepsilon}} \Big/ N_v$$

and CDmean was computed as

$$\sum_{i=1}^{N_v} \frac{c'_i(\mathbf{A}-\lambda(\mathbf{Z'MZ}+\lambda\mathbf{A^{-1}})^{-1})c'_i}{c'_i\mathbf{A}c_i} \Big/ N_v$$

where $\mathbf{M}=\mathbf{I}-\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$. λ was set equal to λ_{Np} according to the size of the TP tested, N_v is the number of validation individuals, and c_i is a contrast vector of length $N_p + N_v$. Our contrast of interest was between the individuals in the validation set and the overall mean of the validation set. For each validation individual *i*, the element in c_i corresponding to the individual *i* was $1-1/N_v$, the elements corresponding to the other validation individuals were $-1/N_v$, and the remaining values were zero. Contrasts were specified in this way because in this case individuals in the TP are not candidates for selection. TPs leading to smaller mean PEVs and larger mean CDs were considered more optimal.

An exchange algorithm was used for the selection of optimal TPs. Step one, a random sample of size *Np* is selected and the optimization criteria of interest is calculated. Step two, a randomly selected individual is removed and then replaced by a new randomly selected individual. Step three, this change is accepted if the TP is improved based on the optimization criteria or rejected if

not. Steps two and three are repeated for a maximum of 2000 iterations or until changes to the TP are rejected for 200 consecutive iterations. The exchange algorithm was repeated 100 times, and the overall optimal TP was selected. GS accuracies using the optimal TPs were computed.

As an external validation, the optimized TPs were used to predict an additional population that was derived by intermating ten individuals selected from the SCs as part of a recurrent selection experiment. Accuracies with optimized TPs from TP_H were compared to accuracies with randomly selected TPs from TP_H. Phenotypic and genotypic data for this external validation population was generated as described for the SC population, except only one season of phenotypic data was available, and mean imputation was used prior to relationship matrix calculation.

Combined TP analysis

TP_{PS} combined with random samples of size *Np* from TP_H was compared to TP_{PS} alone in terms of GS accuracy. Different line mean heritabilities of TP_{PS} and TP_H individuals were simulated. To manipulate the heritability, a random error vector with mean 0 and standard deviation, $\sigma_\varepsilon{}'$ was added to the observations in TP_{PS} and TP_H according to the simulated heritability, H^2_{sim} for both populations:

$$\sigma_\varepsilon{}' = \frac{\sigma_g^2}{H^2_{sim}} - \left(\sigma_g^2 + \frac{\sigma_{ge}^2}{e} + \frac{\sigma_e^2}{er} \right)$$

where σ_g^2, σ_{ge}^2, and σ_e^2 are the genetic, *G*×*E*, and error variances, *e* is the number

of environments and *r* is the number of replicates within an environment. Heritabilities of 0.2 and 0.6 were simulated for TP_{PS} and for each of these heritability levels, *Np* individuals from TP_H were added with H^2_{sim} of 0.2, 0.3, 0.4, or 0.6. GS accuracies were calculated using each combined TP. To determine if accuracy could be improved by weighting the observations from TP_{PS} and TP_H according to the H^2_{sim} of their population of origin, the combined TP analysis was repeated except in the mixed model used for genomic prediction described previously, the diagonal of the residual covariance matrix, *R*, was $1 - H^2_{sim}$.

Results

Phenotypic data characterization

Line mean heritability was 0.82 and 0.61 for the historical and SC populations, respectively. The correlation between the validation set evaluation environments with the historical and population specific training population evaluation environments was 0.81 and 0.83 respectively.

Population characterization

The historical and SC populations were significantly differentiated based on the median Fst across all markers, p-value = 0. Populations also formed distinct but partially overlapping groups together based on their PC1 and PC2 scores (Figure 5.1).

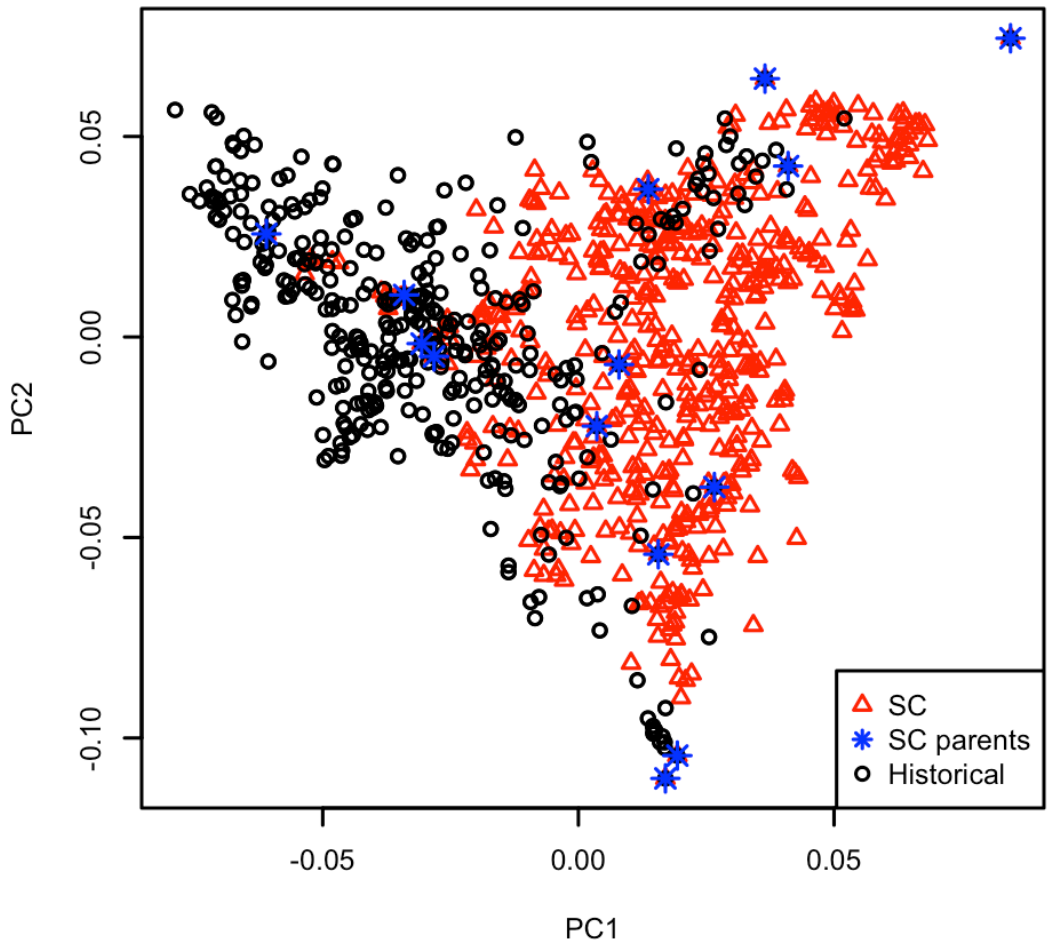


Figure 5.1: Principal components analysis including the historical lines, SCs, and SC parents

The rate of LD decay with physical distance was similar for the historical and SC population, however there was more long-range LD in the SC population (Figure 5.2).

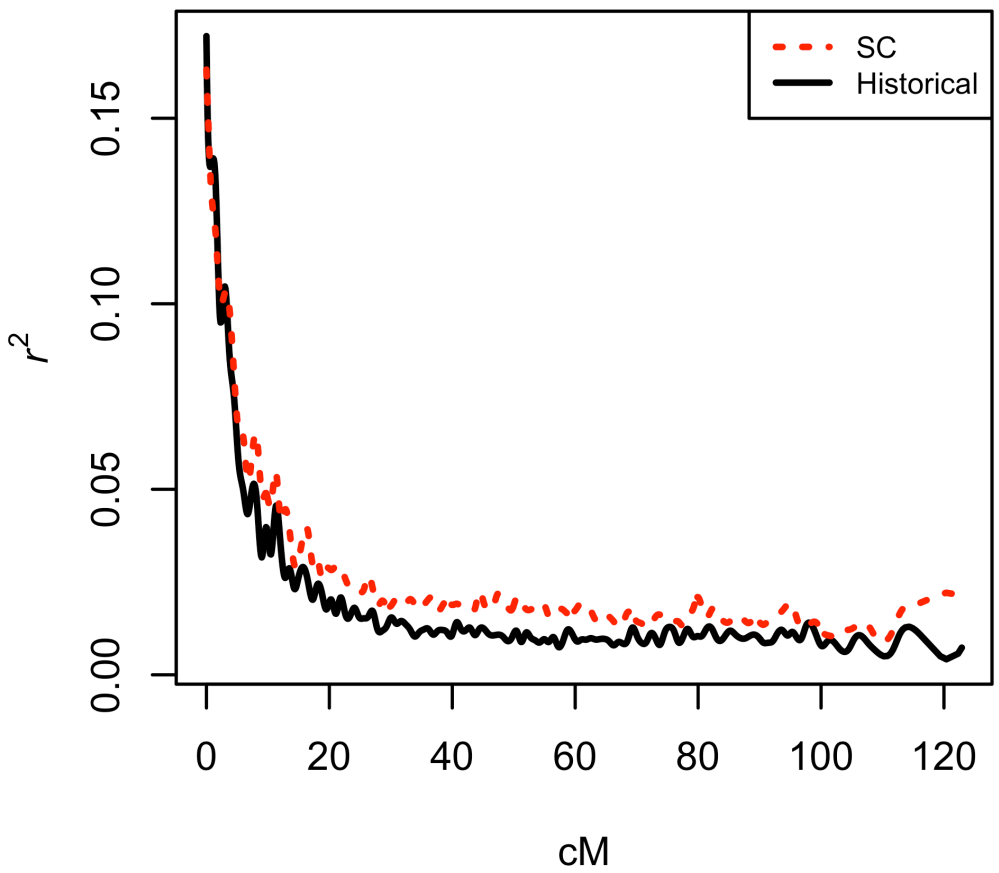


Figure 5.2: Linkage disequilibrium decay with genetic distance within the historical and SC populations

TP comparison and optimization

TP_{PS} always lead to significantly higher accuracies than TP_H and for *Np*= 73 and 146 (Figure 5.3). As *Np* increased, the difference between accuracies from TP_{PS} and TP_H decreased. For example, when *Np* =73, TP_{PS} was 4.4 times more accurate than TP_H, and when *Np* =292, TP_{PS} was only 1.5 times more accurate than TP_H. For TP_H, optimally selected TPs lead to significantly higher accuracies than randomly selected TPs for *Np*= 73, 146, 219, and 292 (Figure 5.4). The optimization criteria PEVmean and CDmean performed similarly and both outperformed Fst. For *Np*= 73, 146, 219, and 292 optimally selected TPs based on

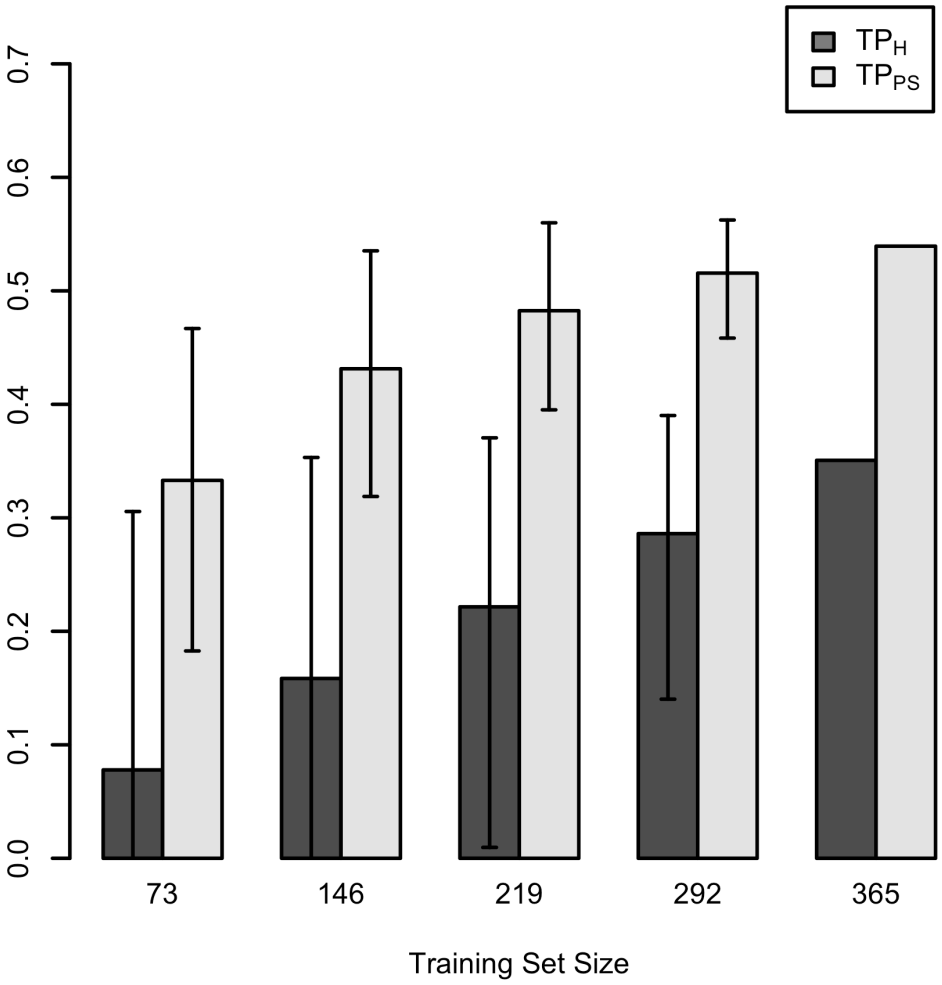


Figure 5.3: Prediction accuracies for the SC population based on TP_{PS} and TP_{H} with varying population sizes

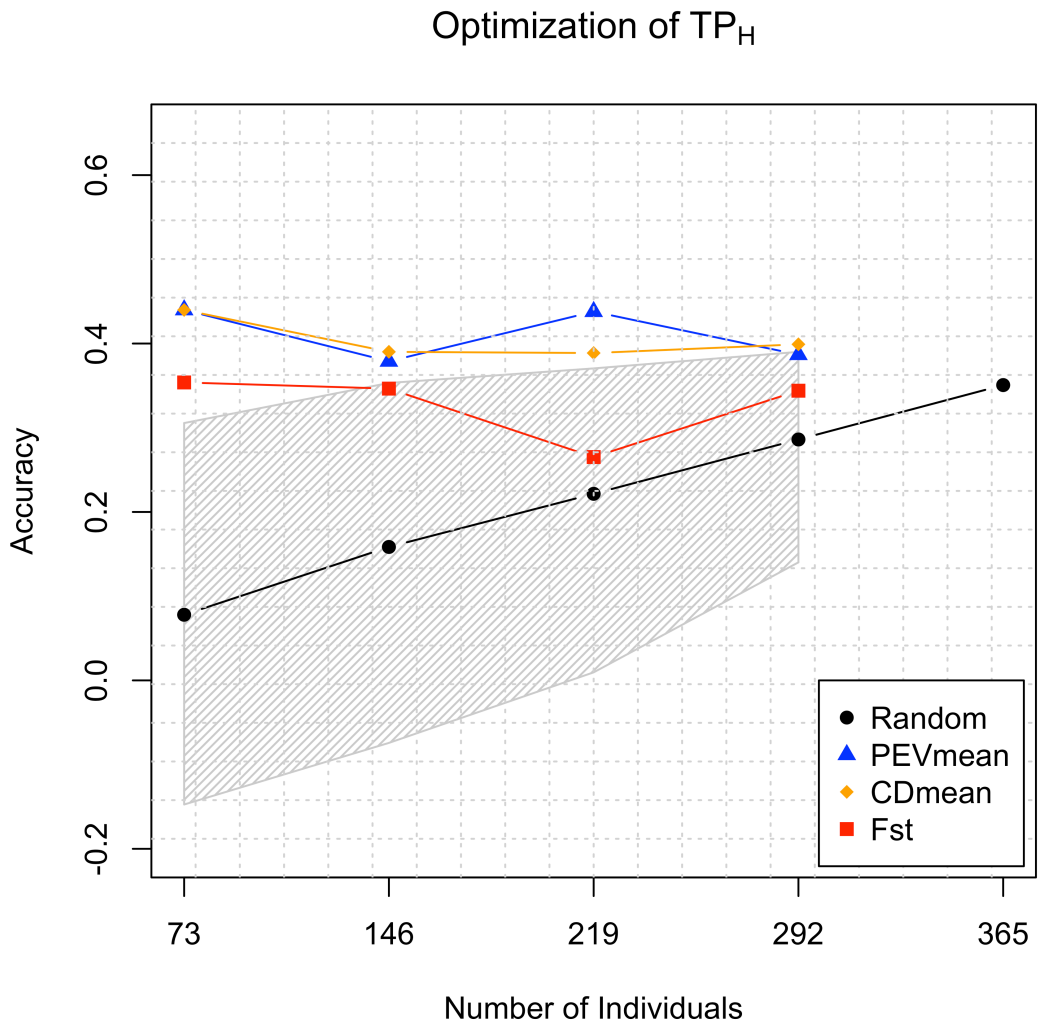


Figure 5.4: Prediction accuracies for the SC population based on optimized TPs from TP_H in comparison with accuracies from randomly sampled TPs from TP_H. The 95% confidence interval for accuracy from randomly sampled TPs is shaded in grey.

PEVmean and CDmean lead to accuracies higher than that of the full TP with *Np*= 365.

When validated using a second population derived from the SC population, the TPs that were optimally selected from TP_H based on CDmean and PEVmean lead to higher accuracies compared to randomly selected TPs (Figure 5).

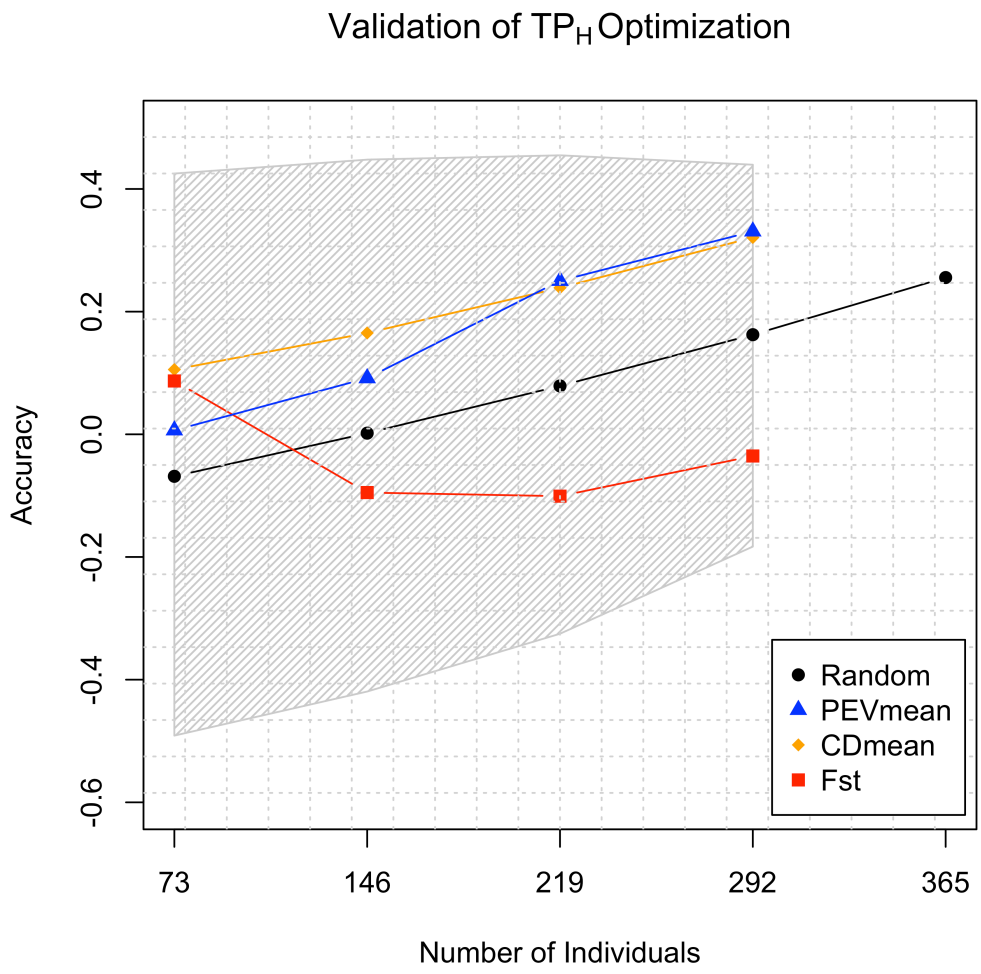


Figure 5.5: Prediction accuracies for an additional validation population based on optimized TPs from TP_H in comparison with accuracies from randomly sampled TPs from TP_H. The 95% confidence interval for accuracy from randomly sampled TPs is shaded in grey.

For this validation experiment, the improvement in accuracy provided by CDmean optimization was most consistent, followed by PEVmean. The TPs selected based on Fst performed worse than random TPs for *Np*= 146, 219, and 292. Although optimized TPs selected using CDmean or PEVmean consistently outperformed random TPs, no significant differences were detected due to the large variation of the random TP accuracies due to sampling. In contrast with the results from validation using the SC population, we observed increasing accuracy as *Np* increased for TPs optimized using CDmean and PEVmean. However, when *Np*=292 and TPs were selected based on PEVmean or CDmean, accuracy was higher than that of the complete TP.

Combined TP analysis

When H^2_{sim} of TP_{PS} was low, 0.2, adding samples from TP_H of H^2_{sim} = 0.3, 0.4, and 0.6, led to a small, but constant improvement in accuracy as *Np* increased (Figure 6). When H^2_{sim} of TP_H was also low, 0.2, adding individuals from TP_H to TP_{PS} led to an initial decrease in accuracy, followed by a slight increase with increasing *Np.* For the maximum number of TP_H samples added, 365, accuracy improved by 1.2%, 7.6%, 10.9%, and 11.9% for H^2_{sim} =0.2, 0.3, 0.4, and 0.6, respectively. Adding a weight of $1 - H^2_{sim}$ to the diagonal of the residual covariance only affected accuracy by up to 1.02% (Figure 6, panel B).

When H^2_{sim} of TP_{PS} was high, 0.6 (Figure 7), adding samples from TP_H of equal heritability led to small and constant increases in accuracy with increasing

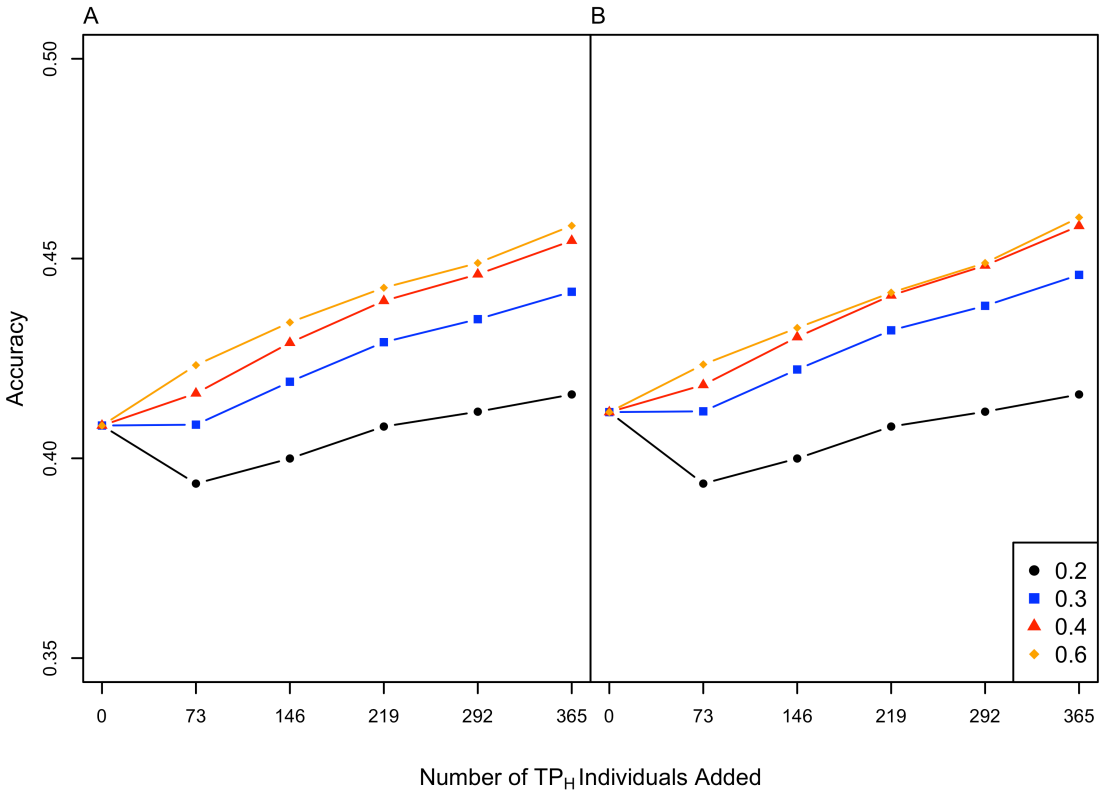


Figure 5.6: The effect of adding TP_H individuals to TP_{PS} when simulated heritability of TP_{PS} is 0.2 and simulated heritability of TP_H is 0.2, 0.3, 0.4, and 0.6. A) Populations are weighted equally, B) populations weighted according to simulated heritability.

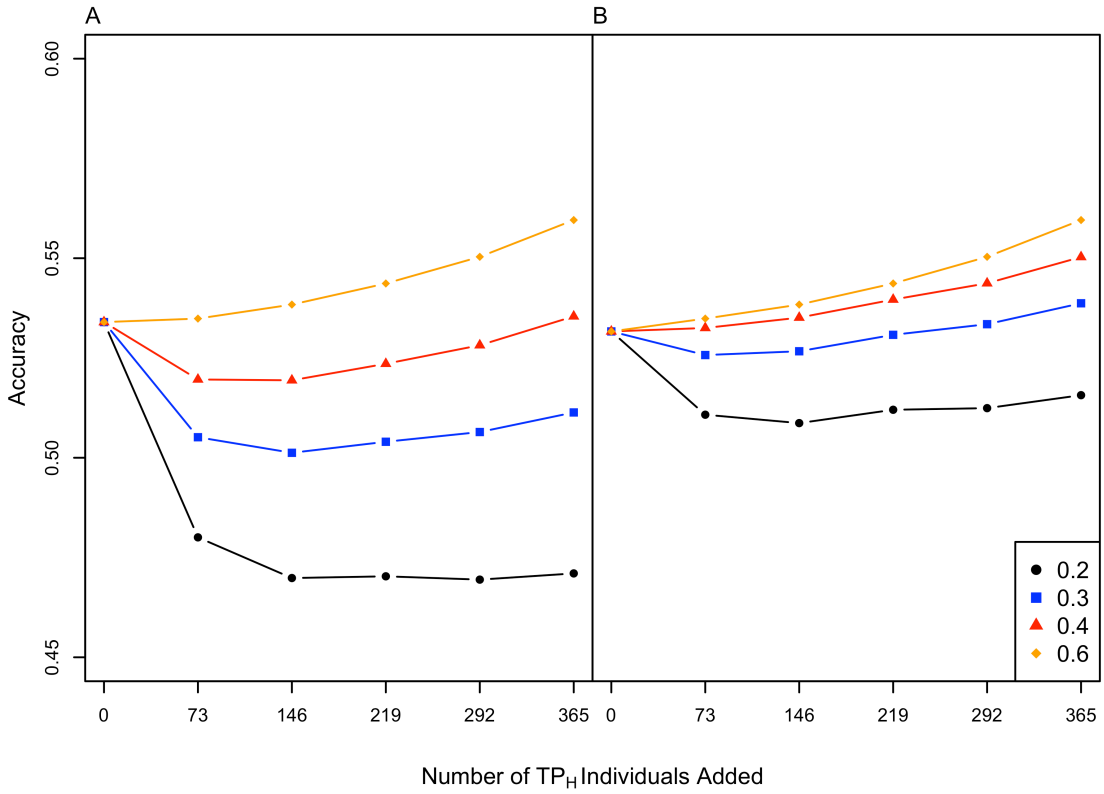


Figure 5.7: The effect of adding TP_H individuals to TP_{PS} when simulated heritability of TP_{PS} is 0.6 and simulated heritability of TP_H is 0.2, 0.3, 0.4, and 0.6. A) Populations are weighted equally, B) populations weighted according to simulated heritability.

Np. When H^2_{sim} of added samples from TP_H was moderate, 0.3 and 0.4, there was an initial decrease in accuracy, followed by a slow increase with increasing *Np*. However, even for the largest *Np*, 365, adding individuals from TP_H led to a -4.5% and 0% change in accuracy for H^2_{sim} = 0.3 and 0.4, respectively. When H^2_{sim} of added samples from TP_H was low, 0.2, accuracy declined by 12% and did not show an eventual increase with increasing *Np*. Adding a weight of $1 - H^2_{sim}$ to the diagonal of the residual led to improved accuracy when H^2_{sim} of TP_H was 0.2, 0.3, and 0.4 (Figure 7, panel B). Improvements in accuracy due to weighting ranged from 2.4% to 9.7%. The lower the H^2_{sim} of TP_H, the greater the benefit of using TP specific weights. However when H^2_{sim} of TP_H was very low, 0.2, adding individuals from TP_H never led to a net increase in accuracy, even when weighing was used.

In summary, using TP_H individuals when TP_{PS} individuals were available for training was always beneficial when H^2_{sim} of TP_H was greater than H^2_{sim} of TP_{PS}. In some cases, when H^2_{sim} of TP_{PS} was high and H^2_{sim} of TP_H was at least moderate, using TP_H and TP_{PS} individuals for training was beneficial when observations were properly weighted according to the heritability of their TP of origin.

Discussion

Populations

The significant population differentiation between the historical and selection candidate populations was a consequence of selection and genetic drift

that occurred because the SC population was generated from only fourteen founder lines from the historical population that were selected because they had at least moderate stem rust resistance and good agronomic performance. Drift could have also occurred during the SC population generation stage, but because Fst between founders and the SC population was very low, 5e-4, drift during this stage was considered negligible. Selection and drift lead to a reduced rate of LD decay with physical distance in the SC population. The level of differentiation between historical and selection candidate populations due to selection and drift would be expected in plant breeding programs because each cycle of selection is founded by a small number of parents selected for intermating. However, breeding programs that use a lower selection intensity may experience less differentiation between historical and selection candidate populations. Thus, the level of selection intensity of a breeding program would have direct implications on the effectiveness of historical data for the prediction of new SCs. Breeding programs with lower selection intensities may be able to use historical data more successfully compared those that use higher selection intensities.

Accuracy comparison

In general, the relative performance of a historical and population specific TP depends on the relative population sizes, heritabilities, levels of $G\times E$, and genetic differentiation between the historical TP and the SCs. Because we observed that TP_H heritability was higher than that of TP_{PS} and GxE between TP_H and the validation set was equal to that of TP_{PS}, it is clear in our case that the

relatively low accuracy from TP_H was primarily driven by the genetic differentiation between TP_H and the validation set. Accuracy from a historical TP could be as high or higher than that of a population specific TP in some scenarios. For example, based on linear regression of accuracy on TP size for both TP_{PS} and TP_H, if we were to add 225 more historical individuals to TP_H, TP_{PS} and TP_H accuracies may have been equivalent. Furthermore, if this study focused on a trait such as yield with low heritability on a single plot basis and high GxE, population specific training data from very few environments will be of low line mean heritability and may not adequately sample the target environments of the breeding program. These factors would lead to a greater advantage of historical data *vs.* population specific data. Thus, this study presents a worst case scenario for the utility of historical data compared to population specific data for this set of germplasm.

Training population optimization

The TP optimization methods based on PEVmean and CDmean enabled the selection of TPs from TP_H that were more accurate than those selected based on random sampling. Other studies evaluating TP optimization (Isidro et al. submitted; Rincent et al. 2014) found similar results; however Isidro et al. (submitted) found that TP optimization could be less accurate than random sampling if it resulted in a reduction in the phenotypic variance. TP optimization would be useful if the historical dataset used for model training does not contain phenotypic data for all traits of interest. A subset of individuals from a historical

dataset, selected to be predictive of the selection candidates based on CDmean or PEVmean, could be phenotyped for new traits of interest. This could reduce costs with potentially little to no sacrifice in accuracy compared to phenotyping all historical individuals.

It may be useful to optimize the training set for future SCs by optimizing with respect to their progenitors (most recent ancestors). When CDmean or PEVmean optimization of the historical TP was done with respect to population progenitors rather than the individuals used for GS validation, accuracy from the optimized TPs was higher than accuracy from random TPs. Although optimizing with respect to population progenitors could be an effective way to select the appropriate subsample individuals for phenotyping and model updating, we could not predict in advance which *Np* would lead to the highest accuracy. Accuracy was maximized when *Np=292*, but in the progenitors accuracy was maximized when *Np=73.* This may have occurred because optimization based on progenitors leads to the selection of sub-optimal TPs. Nevertheless, if phenotyping resources are limited, one could select *Np* based on resource constraints, and select individuals for phenotyping and model updating by optimizing with respect to the progenitors of the future SCs.

Although we observed that for some *Np,* optimal TPs lead to greater accuracy compared to the complete TP, it is not possible to know in advance what *Np* value could maximize accuracy. Furthermore, the ability to select an optimal TP that leads to greater accuracy compared to the complete TP is expected to be

highly dataset and trait dependent due to differences in population and family structure, non-additive genetic variance, and LD between markers and causal loci. Assuming there is perfect linkage between markers and QTL, increasing *Np* values will lead to an asymptotic increase in accuracy (Daetwyler et al., 2010; de Los Campos et al., 2013), and the complete TP will lead to higher accuracy than an optimal subset of the TP. Assuming imperfect linkage between markers and QTL and population or family structure, optimizing the TP for the SCs could increase accuracy because the estimated relationships between pairs of closely related individuals will be more accurate than the estimated relationships between less related individuals (de Los Campos et al., 2013), and eliminating less related individuals could reduce noise in the relationship matrix. The effective population size (N_e) of SC + optimized TP will also be less then that of the SC + full TP, leading to a lower *Np* required for the SC + optimized TP. Aside from population genetic factors, the importance of non-additive genetic variance for the trait of interest may also partially determine if TP optimization improves accuracy. Non-additive genetic variance contributes to the covariance among close relatives only. When the TP is selected to be closely related to the SCs, more non-genetic variance may be captured in G-BLUP, thus for traits where non-additive genetic variance is important, TP optimization may lead to higher accuracy compared to the complete TP. This is similar to the effect of using a Gaussian kernel, where the genetic covariance can decrease more rapidly with genetic distance (Endelman, 2011). However with optimization, the relationship

between some pairs of individuals is effectively set to zero. Because of the various factors that affect the potential gain from training with an optimized TP rather than the complete TP, selecting optimal subsets from a TP is not a reliable way to improve accuracy.

Combining training population data sources

Our results suggest that retaining historical data when data on close relatives are available can reduce accuracy. This was especially pronounced when the heritability of the historical data was low and the heritability of the close relative training data was high. In cases where including historical data was beneficial, the benefit was very small and proper weighting of observations was important. This result has implications for prediction model updating. In a selection program, it may be better to discard older training data that is less relevant to the selection candidates as newer training data becomes available. However, when to discard training data will need to be determined empirically because it will depend on the selection intensity of the breeding program, the availability of data on close relatives, and quality of the historical data. For example, Asoro et al. (2011) evaluated the utility of adding historical oat lines to a training population going back in time and found that historical lines did not decrease accuracy, though the increase in accuracy they provided was small.

We did not test the effect of combining an optimally selected sample of historical data with the population specific data, but we expect that after adding the optimal set of 73, we would observe approximately the same level of accuracy

as we observe when adding the full set of 365, as was observed when we tested historical TP optimization.

Conclusion

This case study found that historical data could be useful for initializing a GS based breeding program where the selection candidates are founded by historical individuals. Although the highest accuracy could be achieved by phenotyping and model training with a subset of the selection candidate population itself, such an approach would require at least two years of additional time to collect multi-location and multi-year data for all traits of interest. While historical data may be useful initially, this study suggests that once GS model updating can occur, it may be best to discard historical data and simply use the most recent data for model training.

Optimization of the historical TP was promising for selection of data subsets that were more predictive than randomly selected subsets. This would be useful when using a historical TP that lacks data for some key traits. To save resources, a subset of the historical TP, rather than the entire TP, could be phenotyped while the selection candidates are being developed.

We note that our conclusions are relevant to the germplasm and trait used in this study, and individual breeding programs will need to initiate GS programs in order to empirically determine the utility of historical data, and at what point data should be discarded from the model training dataset. The utility of TP optimization should also be empirically studied in the context of a GS breeding

program. More publicly available data generated by GS selection experiments and breeding programs will enable many such studies that will lead to the discovery of common trends across datasets.

Acknowledgements

This research was funded by The Bill &Melinda Gates Foundation (Durable Rust Resistance in Wheat) and the United States Department of Agriculture[1]-Agricultural Research Service[1] (USDA-ARS) (Appropriation No. 5430-21000-006-00D). Partial support for J. Rutkoski was provided by a USDA National Needs Fellowship Grant #2008- 38420-04755 and an American Society of Plant Biology (ASPB) -Pioneer Hi-Bred Graduate Student Fellowship.

References

Amin, N., C.M. van Duijn, and Y.S. Aulchenko. 2007. A genomic background based method for association analysis in related individuals. PLoS One 2: e1274.

Asoro, F.G., M. A. Newell, W.D. Beavis, M.P. Scott, Tinker, N. A. and J.-L. Jannink. 2011. Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. Plant Gen. 4: 132–144.

Astle, W., and D.J. Balding. 2009. Population structure and cryptic relatedness in genetic association studies. Stat. Sci. 24: 451–471.

Bates, D., and M. Maechler. 2010. lme4: Linear mixed-effects models using S4 classes. Available at http://cran.r-project.org/package=lme4.

Beeck, C.P., W. A. Cowling, A. B. Smith, and B.R. Cullis. 2010. Analysis of yield and oil from a series of canola breeding trials. Part I. Fitting factor analytic mixed models with pedigree information. Genome 53: 992–1001.

Bernardo, R. 1994. Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci. 34: 20–25.

Box, G.E., and D.R. Cox. 1964. An analysis of transformations. J. R. Stat. Soc. Ser. B 26: 211–252.

Crossa, J., G.D.L. Campos, P. Pérez, D. Gianola, J. Burgueño, J.L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger, and H.-J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713–724.

Daetwyler, H.D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010b. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185: 1021–1031.

Dawson, J.C., J.B. Endelman, N. Heslot, J. Crossa, J. Poland, S. Dreisigacker, Y. Manès, M.E. Sorrells, and J.-L. Jannink. 2013. The use of unbalanced historical data for genomic selection in an international wheat breeding program. F. Crop. Res. 154: 12–22.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: e19379.

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. 4: 250–255.

Garrick, D.J., J.F. Taylor, and R.L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41: 55.

Gilmour, A.R., B.J. Gogel, B.R. Cullis, and R. Thompson. 2009. ASReml user guide release 3.0.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136: 245–257.

Habier, D., R.L. Fernando, and J.C.M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177: 2389–2397.

Haley, C. S., and Visscher, P. M. 1998. Strategies to utilize marker-quantitative trait loci associations. J. Dairy Sci. 81: 85–97.

Hallauer, A.R., M.J. Carena, and J.B. Miranda F. 2010. Quantitative genetics in maize breeding. Iowa State University Press, Ames, IA.

Hayes, B.J., P.M. Visscher, and M.E. Goddard. 2009. Increased accuracy of selection by using the realized relationship matrix. Genet. Res. (Camb). 91: 47–60.

Henderson, C.R. 1984. Applications of linear models in animal breeding. University of Guelph Press, Guelph, Ontario, Canada.

Kennedy, B.W., and D. Trus. 1993. Considerations on genetic connectedness between management units under an animal model. J. Anim. Sci. 71: 2341–2352.

Laloë, D. 1993. Precision and information in linear models of genetic evaluation. Genet. Sel. Evol. 25: 1–20.

Leutenegger, A.-L., B. Prum, E. Génin, C. Verny, A. Lemainque, F. Clerget-Darpoux, and E. A. Thompson. 2003. Estimation of the inbreeding coefficient through use of genomic data. Am. J. Hum. Genet. 73: 516–23.

Long, N., D. Gianola, G.J.M. Rosa, and K.A. Weigel. 2011. Long-term impacts of genome-enabled selection. J. Appl. Genet.: 467–480.

De Los Campos, G., A.I. Vazquez, R. Fernando, Y.C. Klimentidis, and D. Sorensen. 2013. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. 9: e1003608.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Peterson, R.F., A.B. Campbell, and A.E. Hannah. 1948. A diagrammatic scale for estimating rust intensity on leaves and stems of cereals. Can. J. Res. 26c: 496–500.

Piepho, H.P. 2009. Ridge regression and extensions for genomewide selection in maize. Crop Sci. 49: 1165–1176.

Poland, J. A., P.J. Brown, M.E. Sorrells, and J.-L. Jannink. 2012a. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7: e32253.

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sánchez-Villeda, M. Sorrells, and J.-L. Jannink. 2012b. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Gen. 5: 103–113.

Pszczola, M., T. Strabel, H. A. Mulder, and M.P.L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95: 389–400.

R Development Core Team. 2010. R: A language and environment for statistical computing. Available at http://www.r-project.org.

Rincent, R., D. Laloë, S. Nicolas, T. Altmann, D. Brunel, P. Revilla, V.M. Rodríguez, J. Moreno-Gonzalez, A Melchinger, E. Bauer, C.-C. Schoen, N. Meyer, C. Giauffret, C. Bauland, P. Jamin, J. Laborde, H. Monod, P. Flament, A. Charcosset, and L. Moreau. 2012. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). Genetics 192: 715–28.

de Roos, A.P.W., B.J. Hayes, and M.E. Goddard. 2009. Reliability of genomic predictions across multiple populations. Genetics 183: 1545–1553.

Rutkoski, J.E., J. Poland, J.-L. Jannink, and M.E. Sorrells. 2013. Imputation of unordered markers and the impact on genomic selection accuracy. G3 (Bethesda). 3: 427–439.

Storlie, E., and G. Charmet. 2013. Genomic selection accuracy using historical data generated in a wheat breeding program. Plant Gen. 6: 1–9.

Thompson, R., Cullis, B., Smith, A., & Gilmour, A. (2003). A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. Aust. NZ. J. Stat. 45: 445–459.

Wright, S. 1949. The genetical structure of populations. Ann. Eugen. 15: 323–354.

Yu, L.-X., A. Lorenz, J. Rutkoski, R.P. Singh, S. Bhavani, J. Huerta-Espino, and M.E. Sorrells. 2011. Association mapping and gene-gene interaction for stem rust resistance in CIMMYT spring wheat germplasm. Theor. Appl. Genet. 123: 1257–1268.

Zadoks, J., T. Chang, C. F. Konzak, 1974. A decimal code for the growth stages of cereals. Weed. Res., 14:415–421.

CHAPTER 6

GENETIC GAIN FROM PHENOTYPIC AND GENOMIC SELECTION FOR QUANTITATIVE ADULT PLANT STEM RUST RESISTANCE IN WHEAT[6]

Abstract

Stem rust of wheat (*Triticum aestivum* L.), caused by *Puccinia graminis* f.sp. *tritici,* is a globally important disease that can cause complete yield loss. Since the emergence of Ug99, a group of races capable of infecting up to 90% of the worlds' wheat germplasm, breeding for quantitative resistance (QR) has become important for developing varieties with durable, race non-specific resistance. Genomic selection (GS) is a breeding method that could increase rates of genetic gain for quantitative traits such as QR because selection can be based on markers only. Few GS experiments have been conducted in crops, and selection experiments comparing GS based on markers only and phenotypic selection (PS) have not been conducted. Our objectives were to compare realized gain from GS based on markers only with that of PS for stem rust QR; determine if realized gain is consistent with theoretical expectations; and compare the impact of GS and PS on inbreeding, genetic variance, and correlated response for pseudo-black chaff (PBC), a correlated trait. In two years, GS led to a 30.5±10.5% increase in stem rust QR and a 252±35.9% increase in PBC; PS led to a 41.9±11.8%

[6] A paper to be re-formatted and submitted to PNAS as: Rutkoski J. E., J. Poland, R.P. Singh, J. Huerta-Espino, S. Bhavani, J-L. Jannink, M. E. Sorrells. Genetic gain from phenotypic and genomic selection for quantitative adult plant stem rust resistance in wheat. PNAS.

increase in QR and a 276±134% increase in PBC. Selection responses were significant and agreed with theoretical expectations. Loss of genetic variance occurred at a faster rate with GS. These results show that, while GS and PS can lead to similar short-term gains per unit time, GS can lead to faster reductions in genetic variance.

Abbreviations

APR, Adult plant resistance; BLUP, Best linear unbiased prediction; C0, Cycle zero; C1, Cycle one; C1GS-1, Cycle one genomic selection replicate one; C1GS-2, Cycle one genomic selection replicate two; C1PS-1, Cycle one phenotypic selection replicate one; C1PS-2, Cycle one phenotypic selection replicate two; C2, Cycle two; C2GS-1, Cycle two genomic selection replicate one; C2GS-2, Cycle two genomic selection replicate two; GS, Genomic selection; PBC, Pseudo-black chaff; PS, Phenotypic selection; QR, Quantitative resistance.

Introduction

Stem rust of wheat (*Triticum aestivum* L.), caused by the fungal pathogen *Puccinia graminis* f.sp. *tritici,* is a globally widespread and highly damaging disease capable of causing complete yield loss in susceptible varieties (Park, 2007). Although major epidemics of stem rust have not been recorded since the 1950s, in 1998 a new race group, Ug99, capable of infecting over 90% of the world's wheat germplasm (Singh et al., 2008) was discovered in Uganda. Ug99 has since been migrating via wind currents, reaching as far as South Africa (Pretorius et al., 2010) and Iran (Nazari et al., 2009). As the pathogen spreads, it

continues to evolve, overcoming an even larger set of major-effect resistance genes (Jin et al., 2008, 2009). The emergence and continued evolution of Ug99 has prompted efforts to rapidly develop Ug99 resistant varieties adapted to vulnerable regions.

Resistance to stem rust can be either qualitative or quantitative. The qualitative form of resistance is based on race-specific pathogen recognition genes (R-genes) that interact with the pathogen in a gene-for-gene manner (Flor, 1971). Although a single R-gene can condition complete resistance against its corresponding races; migration, mutation, and/or selection within pathogen populations can render an R-gene ineffective in a relatively short period of time (McDonald and Linde, 2002). This occurs in regions such as East Africa where environmental conditions are favorable for stem rust pathogen evolution. In these regions, the deployment of quantitative resistance (QR), also referred to as slow-rusting adult plant resistance, is advocated because it is generally durable and non-race specific (Parlevliet, 2002). Like other quantitative traits, stem rust QR is based on multiple small effect loci (Knott, 1982; Singh et al., 2013), and improvement of QR to desirable levels in breeding populations requires multiple cycles of selection.

Pseudo-black chaff (PBC), black discoloration on the glumes and stems, is associated with stem rust QR. At least four loci are involved in PBC expression including the *Sr2* locus (Hare and McIntosh, 1979; Bariana et al., 2001; Yu et al., 2011; Singh et al., 2013), which is associated with both PBC and durable stem

rust QR. Although PBC can be a useful morphological marker for partial stem rust resistance, it is an undesirable trait because it can be misidentified as disease in a farmer's field.

Genomic selection (GS) has been proposed as an appropriate marker-assisted breeding strategy for stem rust QR (Rutkoski et al., 2010). With GS, reviewed by Heffner et al. (2009) and Lorenz et al. (2011), a statistical model trained with phenotypic and genotypic data from a relevant population is used to predict, based on markers only, the breeding values of new selection candidates that have been genotyped. This enables selection to occur before phenotyping, potentially leading to greater genetic gain per unit time. In spring wheat, GS cross-validation studies have been conducted for various traits (Crossa et al., 2010; Poland et al., 2012b; Dawson et al., 2013) including stem rust resistance (Ornella et al., 2012; Rutkoski et al., in press).

Selection experiments are useful for comparing breeding methods, assessing the response of a particular trait to selection, identifying what unselected traits may exhibit a correlated response to selection, and measuring changes in inbreeding and genetic variability due to selection. Selection experiments to evaluate GS have been conducted in maize (*Zea mays* L.) (Massman et al., 2013) and in oats (*Avena sativa* L.) (Asoro et al., 2013). Massman et al. (2013) compared GS with marker assisted recurrent selection for an index of grain yield and stover quality in maize, and found 14 to 50% greater gain from GS. Asoro et al. (2013) compared marker-assisted selection based on markers

and phenotype, pedigree best linear unbiased prediction (BLUP), and GS based on phenotypic data and a marker based relationship matrix for the improvement of ß-glucan in oats and found that marker based methods were significantly more effective than pedigree BLUP. This study also found that selections based on pedigree BLUP tended to be more closely related compared to selections based on GS.

While these studies are informative, a realized gain experiment comparing GS based on markers only with phenotypic selection (PS) selection on a per unit time basis in crop plants has not been done. This is important because in many crops such as wheat, PS is the most commonly used breeding method for quantitative traits. Furthermore, selection studies of GS in wheat or for quantitative disease resistance have yet to be conducted. The objective of this study was to compare realized genetic gain per unit time from GS based on markers only *vs.* PS for stem rust QR in spring wheat, determine if realized gain is consistent with expected gain based on theory, and compare GS with PS in terms of impact on inbreeding and genetic variance and correlated response for PBC.

Materials and methods

Genetic material

Historical population: Three hundred seventy four individuals were selected from the CIMMYT stem rust screening nurseries. Individuals that were suspected to contain race-specific genes effective against Ug99 based on pedigree and adult plant infection type were avoided. As described by Rutkoski et al.

(2013), the absence of major resistance genes effective against stem rust race TTKST was confirmed in 365 of the individuals based on seedling tests conducted at the cereal disease lab, St. Paul, Minnesota.

Cycle zero population: Cycle zero (C0) was derived from 14 individuals (Table 6.1) selected from the historical population based on their agronomic performance, complementarity for different traits, and absence of major resistance genes effective against TTKST. Stem rust QR among the 14 founders ranged from moderately resistant to moderately susceptible. To generate C0, the founders were intermated pseudo-randomly for two generations by hand pollination. For the first round of intermating, a partial diallel crossing scheme (Kempthorne and Curnow, 1961), with each parent involved in seven cross combinations, was used to generate 49 F1s. F1s were confirmed based on SSR genotyping. In the next round of intermating, F1s were intercrossed so that each participated in at least one cross and crosses between F1s with common parents were avoided. 84 successful double cross F1s resulted. Double cross F1s were selfed to increase seed, resulting in double cross F2s. Five hundred four double cross F2 individuals sampled from each of the 84 families became C0. In order to measure the variance in selection response, the 504 individuals were split into two replicate C0 populations, C0-1 and C0-2 of size 253 and 252,respectively.

Genotypic data

Genotypic data was generated in two batches, the first prior to cycle one of selection and included the historical and C0 populations, the second prior to cycle

Table 6.1: C0 founder identifying information

Cornell ID	Cross name	CIMMYT selection history
H2-18	BAJ	CGSS01Y00134S-099Y-099M-099M-13Y-0B
H2-19	KACHU	CMSS97M03912T-040Y-020Y-030M-020Y-040M-4Y-3M-0Y
M5-12	MARCHOUCH*4/SAADA/3/2*FRET2/KUKUNA//FRET2	CGSS05Y00206T-099M-099Y-099M-099Y-099ZTM-7WGY-0B
M5-152	PBW343*2/KHVAKI//PARUS/3/PBW343/PASTOR	CGSS05B00271T-099TOPY-099M-099NJ-099NJ-12WGY-0B
M5-131	PBW343*2/KUKUNA//PARUS/3/PBW343*2/KUKUNA	CGSS05B00256T-099TOPY-099M-099NJ-099NJ-5WGY-0B
M5-147	PBW343*2/KUKUNA//SRTU/3/PBW343*2/KHVAKI	CGSS05B00261T-099TOPY-099M-099NJ-099NJ-8WGY-0B
H2-20	PFAU/SERI.1B//AMAD/3/WAXWING	CGSS02Y00153S-099M-099Y-099M-46Y-0B
H2-53	PICAFLOR #2	CGSS02Y00152S-099M-099Y-099M-11WGY-0B
M5-102	SERI.1B//KAUZ/HEVO/3/AMAD*2/4/KIRITATI	CGSS05B00198T-099TOPY-099M-099NJ-14WGY-0B
M5-43	TACUPETO F2001/BRAMBLING//PVN	CMSS05B00218S-099Y-099M-099NJ-4WGY-0B
M5-100	TRCH/SRTU/5/KAUZ//ALTAR84/AOS/3/MILAN/KAUZ/4/HUITES	CGSS05B00189T-099TOPY-099M-099NJ-099NJ-7WGY-0B
H2-34	WAXWING*2/KIRITATI	CGSS01B00054T-099Y-099M-099M-099Y-099M-13Y-0B
M5-18	WAXWING/6/PVN//CAR422/ANA/5/BOW/CROW//BUC/PVN/3/YR/4/TRAP#1	CGSS05Y00363S-0B-099Y-099M-099NJ-099NJ-6WGY-0B
M5-37	WBLL1*2/CHAPIO//MESIA	CMSS05B00063S-099Y-099M-099Y-099ZTM-6WGY-0B

two of selection and included the cycle one (C1) population. All genotypic data was generated using genotyping-by-sequencing (GBS, Elshire et al., 2011) according to the protocol described in Poland et al. (2012). Due to the heterozygosity in the C0 and C1 populations, polymorphisms were called in C0 and C1 based on the polymorphic markers that were discovered in the historical population. In total there were 27,434 markers.

Marker re-coding and filtering was also carried out in two batches. Batch one contained the historical and C0 populations, and batch two contained the historical, C0, and GS C1 populations. Marker genotypes were recoded as -1, 0, 1. Homozygotes for the minor allele were coded as -1, heterozygotes were coded as 0 and homozygotes for the major allele were coded as 1. For the first batch of genotypic data, markers with more than 80% missing data were removed. For the second batch of genotypic data, non-redundant markers with pairwise $r^2 < 1$ were selected and markers with 80% missing data or more were removed. This resulted in 20,882 and 18,653 markers remaining after marker editing in the first and second batch respectively. Mean imputation was used to handle missing data. In the C0 and C1 populations, mean imputation was carried out within full-sib families.

Phenotypic data

Stem rust QR was phenotyped at the international Ug99 stem rust screening nurseries at the Kenya Agricultural Research Institute, Njoro, Kenya and the Ethiopian Institute of Agricultural Research, Debre Zeit, Ethiopia as

described in Rutkoski et al. (2013). The training population, consisting of the historical population, was evaluated across 18 environments between 2005 and 2012, with each individual appearing in about four environments. New populations were evaluated as soon as seed was available. C0 populations were evaluated in the Njoro 2012 off-season, Njoro 2012 main-season, Njoro 2013 off-season, and Debre Zeit 2013 off-season. C1 populations were evaluated in Njoro 2012 main-season, Njoro 2013 off-season, and Debre Zeit 2013 off-season. Cycle two (C2) populations were evaluated in the Njoro 2014 off-season, and Debre Zeit 2014 off-season. An augmented lattice square design (Federer, 2002) was used for the C0, C1, and C2 population trials. Checks consisted of a sample of historical individuals and C0, C1, or C2 individuals with abundant seed. Phenotypic data from 2012 and 2013 was Box-Cox (Box and Cox, 1964) transformed prior to all analyses to avoid non-normal residuals.

Genomic selection cycle one

GS model training was done in two stages. In the first stage, genetic values of the historical individuals were estimated using phenotypic data collected between 2005 and 2011 (TrainingPhenoData1.csv). The R (R Development Core Team, 2010) package *lme4* (Bates & Maechler, 2010) was used to fit the mixed model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$ where $\mathbf{Y}$ is a vector of phenotypes, $\boldsymbol{\beta}$ is a vector of environment effects treated as fixed, and $\mathbf{u}$ is a vector of genotype effects treated as random. $\mathbf{X}$ and $\mathbf{Z}$ were the design matrices relating observations to environments and genotypes. The solutions for $\mathbf{u}$ were used as the genetic values

$\mathbf{Y_{GV}}$ of the historical population.

In the second stage, $\mathbf{Y_{GV}}$ of the historical individuals were used in Bayesian ridge regression (Pérez, de Los Campos, Crossa, & Gianola, 2010) based on the general model, $\mathbf{Y_{GV} = Mx}$. Where $\mathbf{Y_{GV}}$ is a vector of genetic values, **M** is the marker genotype matrix and **x** is a vector of marker effects. Bayesian ridge regression assumes Gaussian distributed marker effects with common marker variances and is the Bayesian equivalent to ridge regression BLUP. Predicted breeding values of the C0 populations were estimated as $\mathbf{Mx}$

The five individuals with the highest predicted breeding values for stem rust resistance were selected for intermating. For each replicate C0 population, C0-1 and C0-2, the five selections were intermated based on their S1 progeny using a half-diallel crossing scheme. At least six S1 progenies were used for each selected individual. Two to three successful crosses were made per combination. These pseudo-F1s were selfed for seed increase, creating the cycle one GS replicate one (C1GS-1) and cycle one GS replicate two (C1GS-2) populations.

Phenotypic selection cycle one

The historical and C0 genetic values were calculated using phenotypic data collected between 2005 and 2012 (TrainingPhenoData2.csv). ASReml-R (Gilmour et al., 2009) was used to fit a mixed model with a fixed environment effects and random genotype effects. A random block effect was modeled in the Kenya main-season 2012 environment. A separable first order autoregressive variance model was used to model the variance structure of the plot errors in the

row and column direction within the Kenya off-season 2012 environment. The covariance among genotypes was modeled as proportional to a pedigree relationship matrix (Ped.csv). This model was selected based on AIC and BIC. The solutions for the random genotype effect were used as the genetic values of the historical and C0 individuals.

The five C0 individuals with the highest breeding values based on the mixed model analysis were selected from C0-1 and C0-2. Intermating and selfing were carried out in GS cycle one, creating the cycle one PS replicate one (C1PS-1) and cycle one PS replicate two (C1PS-2) populations.

Genomic selection cycle two

Due to computational advantages, predictions of the C1 individuals were generated using a genomic BLUP model implemented in the R package *rrBLUP* (Endelman, 2011). The genomic BLUP model is $\mathbf{Y_{GV}} = \mathbf{Zu} + \boldsymbol{\varepsilon}$, $\mathbf{u} \sim \mathrm{N}(0, \mathbf{G}\sigma_u^2)$. Where $\mathbf{Y_{GV}}$ is a vector of the genetic values of the historical and C0 population that were calculated previously for PS cycle one, $\mathbf{G = MM'}$ is a marker based relationship matrix containing the historical, C0 and C1 individuals. The solutions for *u* contained the predicted breeding values for the C1 individuals. Breeding value predictions from genomic BLUP with $\mathbf{G = MM'}$ are equivalent to those from ridge regression BLUP (Hayes et al., 2009).

Within both C1GS-1 and C1GS-1, the five individuals the highest predicted breeding values for stem rust resistance were selected for intermating. Intermating and selfing were carried out as in GS cycle one, the cycle two GS

replicate one (C2GS-1) and cycle two GS replicate two (C2GS-2) populations.

Expected gain from selection for stem rust quantitative resistance

Expected gain from selection per cycle for PS was calculated using the general formula $\Delta G = \sum_i k_i c_i \hat{\sigma}_A^2 / \hat{\sigma}_{yi}$ (Hallauer et al., 2010) with *i* corresponding to different sexes or different selection units, and where $\hat{\sigma}_A^2$, is the additive genetic variance, *k* is the selection intensity, *c* is the covariance between the selection units and the individuals in the improved population, $\hat{\sigma}_y$ is the phenotypic standard deviation on a line mean basis, and $k = (\mu_s - \mu) / \hat{\sigma}_y$, where μ_s is the mean of the selected individuals and μ is the population mean.

$\hat{\sigma}_y = \sqrt{\hat{\sigma}_g^2 + \hat{\sigma}_{ge}^2 / e + \hat{\sigma}_\varepsilon^2 / er}$, where $\hat{\sigma}_g^2$ is the genetic variance, $\hat{\sigma}_{ge}^2$ is the genotype by environment interaction variance, $\hat{\sigma}_\varepsilon^2$ is the error variance, *e* is the number of environments, and *r* is the number of replicates within environment. For GS, the expected gain from selection per cycle was calculated based on the general formula for gain from correlated trait selection $\Delta G_Y = \sum_i k_{X_i} c_{X_i} h_{X_i} r_{XY_i} \hat{\sigma}_{A_{Y_i}}$, with *Y* corresponding to the trait of interest, and *X* corresponding to the trait directly under selection and where h_X is the selection accuracy of trait *X*, and r_{XY} is the correlation between trait *X* an *Y*. In the case of GS, we assumed h_X=1, and r_{XY} is the GS prediction accuracy. For both PS and GS, selection equation parameters were equal for both sexes. In all cases *c* was equal to 1.5.

To estimate the expected ΔG from the first cycle of PS and GS, variance components, and selection intensities were estimated using Box-Cox transformed (Box and Cox, 1964) phenotypic data for the C0 population recorded in Njoro during the 2012 main and off-seasons. GS accuracy was estimated as the correlation between the C0 predicted breeding values based on the GS model and the estimated breeding values based on phenotype. The estimate of expected ΔG from the second cycle of GS was based on variance component and GS accuracy estimates from Box-Cox transformed (Box and Cox, 1964) phenotypic data for the C1 population recorded in Njoro during the 2013 main season. Expected ΔG was estimated separately for each replicate of GS and PS and were converted back to a non-transformed scale. To express ΔG in terms of gain in resistance levels, ΔG was multiplied by negative one.

Expected correlated response for pseudo-black chaff

Correlated response to selection for PBC was calculated using the formula for gain from correlated trait selection described above, but in this case trait Y was PBC and trait X was either stem rust severity or genomic predictions of stem rust severity. To estimate ΔG from first cycle of GS and PS, $\hat{\sigma}_A$, and r_{XY} were estimated using Box-Cox transformed (Box and Cox, 1964) C0 data collected in Njoro during the 2012 main and off-seasons. For GS, r_{XY} was the correlation between PBC and genomic estimated breeding values for stem rust severity, and for PS, r_{XY} was the correlation between PBC and breeding values for stem rust estimated using phenotype and pedigree. For response from the second cycle of

GS, additive genetic variance and correlations were calculated using C1 data recorded in Njoro during the 2013 main season. For correlated response from PS h_X was 0.82 and 0.76 and r_{XY} was -0.463 and -0.384 for replicates one and two respectively. For GS cycle one replicates one and two, r_{XY} was -0.253, and -0.125; and for cycle two replicates one and two, r_{XY} was -0.234 and -0.381. Expected ΔG values were converted to a non-transformed scale when necessary.

Realized gain from selection for stem rust quantitative resistance

Stem rust severity was evaluated for all populations, C0-1, C0-2, C1PS-1, C1PS-2, C1GS-1, C1GS-2, C2GS-1, and C2GS-2, in Njoro, and Debre Zeit during the 2014 off-season using all individuals with sufficient seed. Number of individuals evaluated per population is shown in Table 6.2.

Table 6.2: Mean stem rust severity, mean PBC, mean level of inbreeding, and genetic variance for each population.

Population†	Number of individuals evaluated	Mean stem rust severity	Mean PBC	Mean level of inbreeding	Genetic variance
C0-1	240	39.1	0.265	0.502	79.9
C1PS-1	94	18.1	1.35	0.57	81.7
C1GS-1	258	35	0.693	0.571	70.4
C2GS-1	288	31.3	0.836	0.695	33.3
C0-2	241	37.8	0.241	0.503	74.4
C1PS-2	241	26.4	0.584	0.527	102
C1GS-2	267	33.7	0.354	0.586	76.1
C2GS-2	280	22.3	0.935	0.734	51.1

†C0-1, cycle zero replicate one; C1PS-1, cycle one PS replicate one; C1GS-1, cycle one GS replicate one; C2GS-1, cycle two GS replicate one; C0-2, cycle zero replicate two; C1PS-2, cycle one PS replicate two; C1GS-2, cycle one GS replicate two; C2GS-2, cycle two GS replicate two

Adjusted population means were estimated using a mixed model with a fixed population effect, and a random environment, and genotype effect and a random block effect within the Njoro environment. A first order autoregressive variance model was used to model the variance structure of the plot errors in the row and column direction. For summary data, this model, excluding the treatment effect, was fit to calculate genetic values for the individuals. To calculate genetic values within environment, the environment effect was also removed, and the model was fit separately for Njoro and Debre Zeit.

Percent total gain, was estimated as $(C_n - C_0)/C_0 \times 100$, where C_n is the mean of the population generated from n cycles of selection, and C_0 is the mean of the base population. To test for significance of selection response, according to Hallauer et al. (2010) the adjusted population means were used in the linear regression model, $P_{ijk} = b_k + \sum_j \beta_j x_{ij} + \varepsilon_{ijk}$, where P_{ijk} is the population mean for replicate k of cycle i of selection method j, b_k is the base population mean for replicate k, and β_j is the rate of selection gain per cycle for selection method j. Percent gain per cycle was calculated as $\beta_j/C_0 \times 100$. Realized ΔG was calculated as $P_{ijk} - b_k$. Paired two tailed t-tests were used to test for differences in ΔG per cycle and per unit time between GS and PS and to test for differences between observed and expected ΔG. To express gain from selection in terms of stem rust resistance, percent total gain, β_j, and realized ΔG were multiplied by negative one.

Mean level of inbreeding and genetic variance

For each set of individuals selected at each cycle of selection, the mean level of inbreeding resulting from a generation of random mating and one generation of selfing was calculated based on pedigrees using the R package *pedigreemm* (Vazquez et al., 2010), which uses the algorithm described in the appendix of Sargolzaei and Iwaisaki (2005) to calculate inbreeding coefficients. The expected inbreeding rate was calculated as $\Delta F = 1/2N_e$ where N_e=5, the effective population size. Actual ΔF per cycle was calculated using the linear regression model $F_{ijk} = 0.5 + \sum_j \delta_j x_{ij} + \varepsilon_{ijk}$, where F_{ijk} is the mean level of inbreeding for replicate *k* of cycle *i* of selection method *j*, 0.5 is the inbreeding coefficient of the base population, and δ_j is the inbreeding rate for selection method *j.*

Genetic variance was estimated for each population using data from Njoro, and Debre Zeit during the 2014 off-season using a two-stage analysis. In the first stage the model included a fixed environment effect, a random block effect within the Njoro environment, and a first order autoregressive variance model for variance structure of the plot errors in the row and column direction. In the second stage, genetic variance for a given population was estimated using only individuals from that population. For both inbreeding and genetic variance, paired two tailed t-tests were used to test for differences between GS and PS per cycle and per unit time.

Correlated response for pseudo-black chaff

To calculate correlated response for PBC, PBC on the glumes was scored on a zero to five scale in Njoro 2014, where conditions were favorable for PBC expression. A Box-Cox (Box and Cox, 1964) transformation was applied prior to analysis. Adjusted population means were estimated using a mixed model with a fixed population effect, and a random genotype effect. Means were then converted back to a non-transformed scale and used to test for significant selection response and to calculate percent gain as previously described. Paired two tailed t-tests were used to test for differences between GS and PS per cycle and per unit time.

Results

Selection cycle duration

The cycle durations were one year for GS and two years for PS (Figure 6.1). Because routine phenotyping for Ug99 resistance in bread wheat only takes place in Njoro during March and September, the exact length of the PS cycle, as well as the number of seasons of data that can be used for GS model updating prior to selection cycle two, depends on the month when the selection scheme is initiated (Table 6.3). Across all possible starting months, assuming that PS is based on two seasons of data, the PS cycle duration ranges from 1.83 to 2.27 years and is 2.05 years on average. For GS, two seasons of data for model updating can occur for all but two starting months.

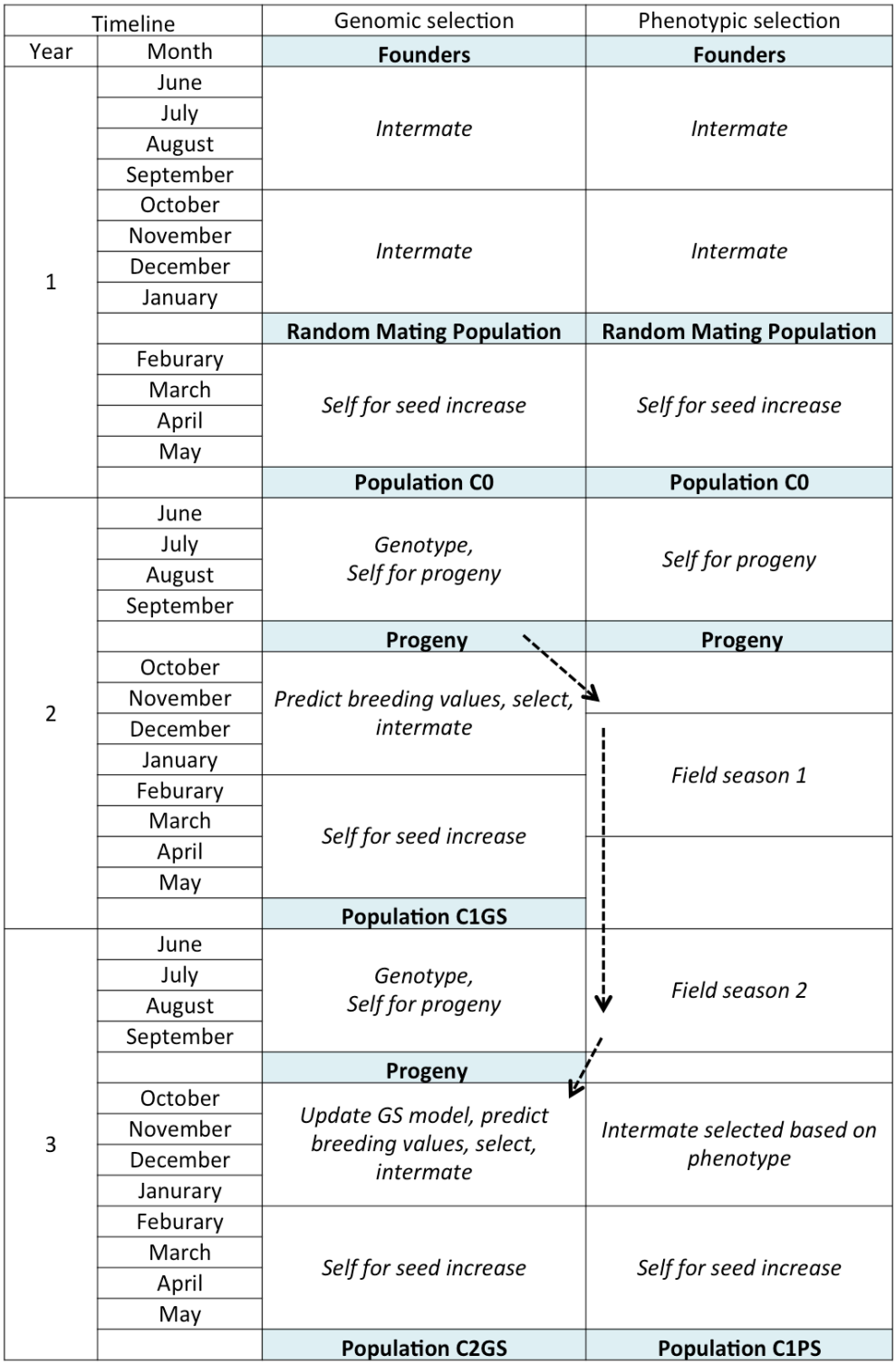


Figure 6.1: Timeline of GS and PS selection schemes for a one year GS cycle and a two year PS cycle. Year one consists of C0 population development and is not part of the breeding cycle. In the genomic selection pipeline, arrows branching from the main pipeline show activities for model updating that occur simultaneously.

Table 6.3: Cycle time for PS and GS, and number of seasons of data that can be used for GS model updating for different starting months

Starting Month	PS cycle time†	GS cycle time	Seasons of data for GS model updating‡
January	1.92	1	2
February	1.83	1	2
March	2.25	1	1
April	2.27	1	2
May	2.08	1	2
June	2	1	2
July	1.92	1	2
August	1.83	1	2
September	2.25	1	2
October	2.17	1	1
November	2.08	1	2
December	2	1	2

† Time is expressed in years
‡ Phenotying seasons are assumed to be December through March, and June through September

Realized gain trial

Environmental conditions at both Njoro and Debre Zeit were exceptionally favorable for stem rust disease development during the 2014 realized gain trial. Disease pressure was slightly higher in Debre Zeit, where mean severity was 36, compared to Njoro, where mean severity was 27. Repeatability within Njoro and Debre Zeit was 0.793 and 0.621 respectively, and line mean heritability across both environments was 0.777. Estimates of genetic values were consistent across both environments (Figure 6.2), with a correlation of 0.665. Population mean by environment interaction was observed (Figure 6.3), however there was only one instance of a crossover interaction. Overall population means for stem rust QR and PBC are summarized in Table 6.2.

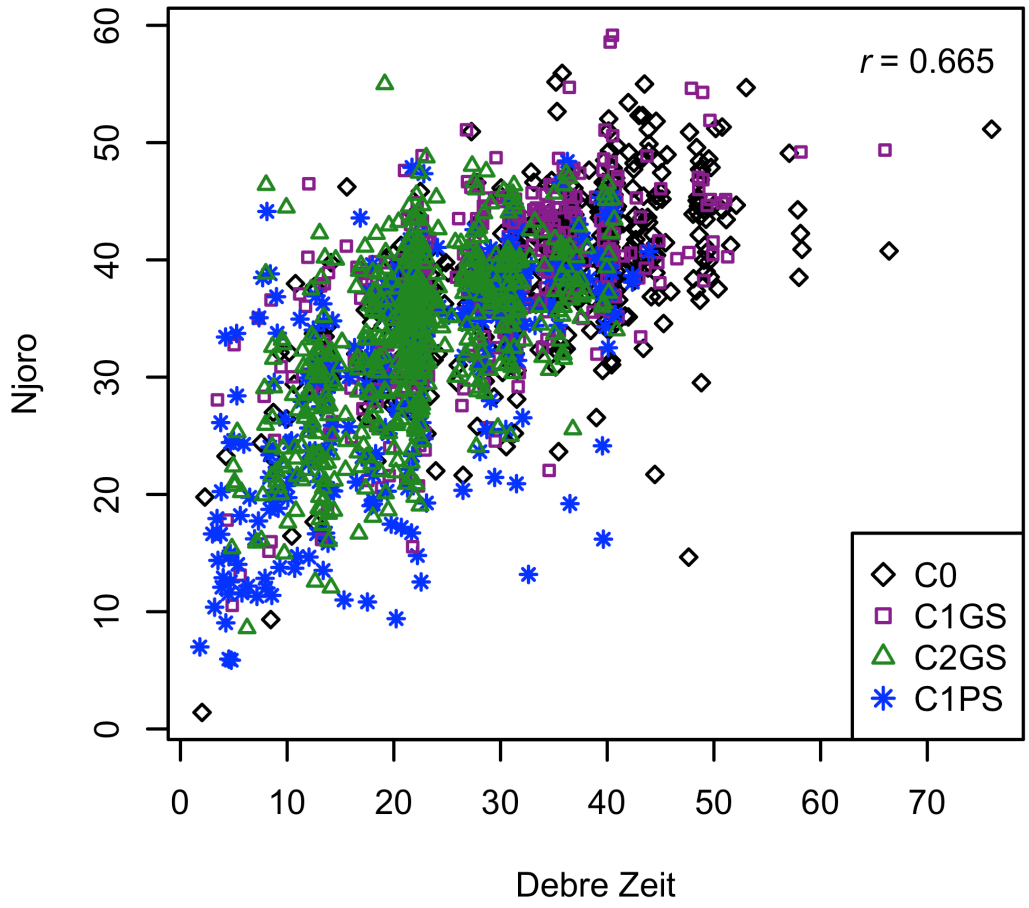


Figure 6.2: Stem rust severity in Njoro *vs.* stem rust severity in Debre Zeit during the 2014 realized gain trial. Correlation between the two environments was 0.66. C0, C1GS, C2GS, and C1PS populations are coded as black diamonds, purple squares, green triangles, and blue stars, respectively.

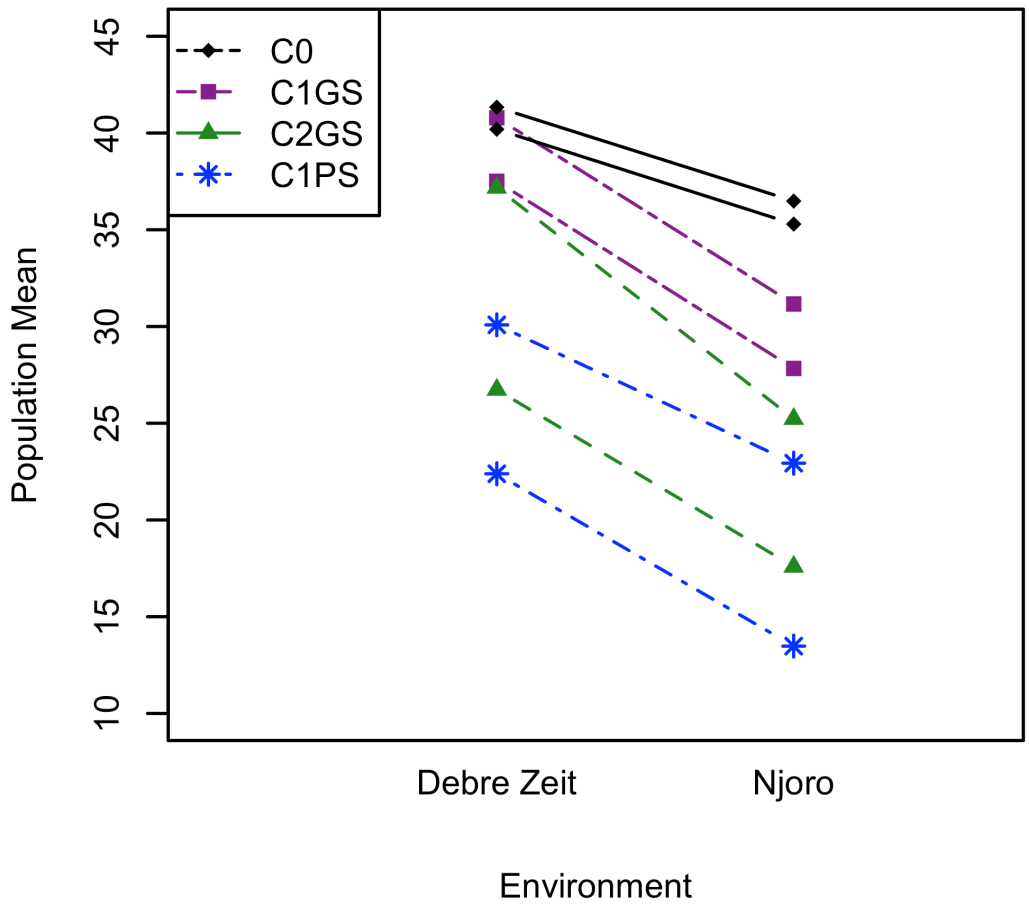


Figure 6.3: Population means across the Debre Zeit and Njoro environments plotted to show population by environment interaction. C0, C1GS, C2GS, and C1PS populations are coded as black diamonds, purple squares, green triangles, and blue stars respectively.

Gain from selection for stem rust quantitative resistance

Mean stem rust severity of C0, C1PS, C1GS, and C2GS was, 38.5±0.664, 22.3±4.15, 34.4±0.677, 26.8±4.5 Expected and realized ΔG for stem rust resistance, expressed as the reduction in severity, was 9.66±3.12 and 16.2±4.82 for PS cycle one, 4.13±0.911 and 4.125±0.0131 for GS cycle one, and 9.55±0.155 and 11.7±3.83 for GS cycle two (Figure 6.4A-B).

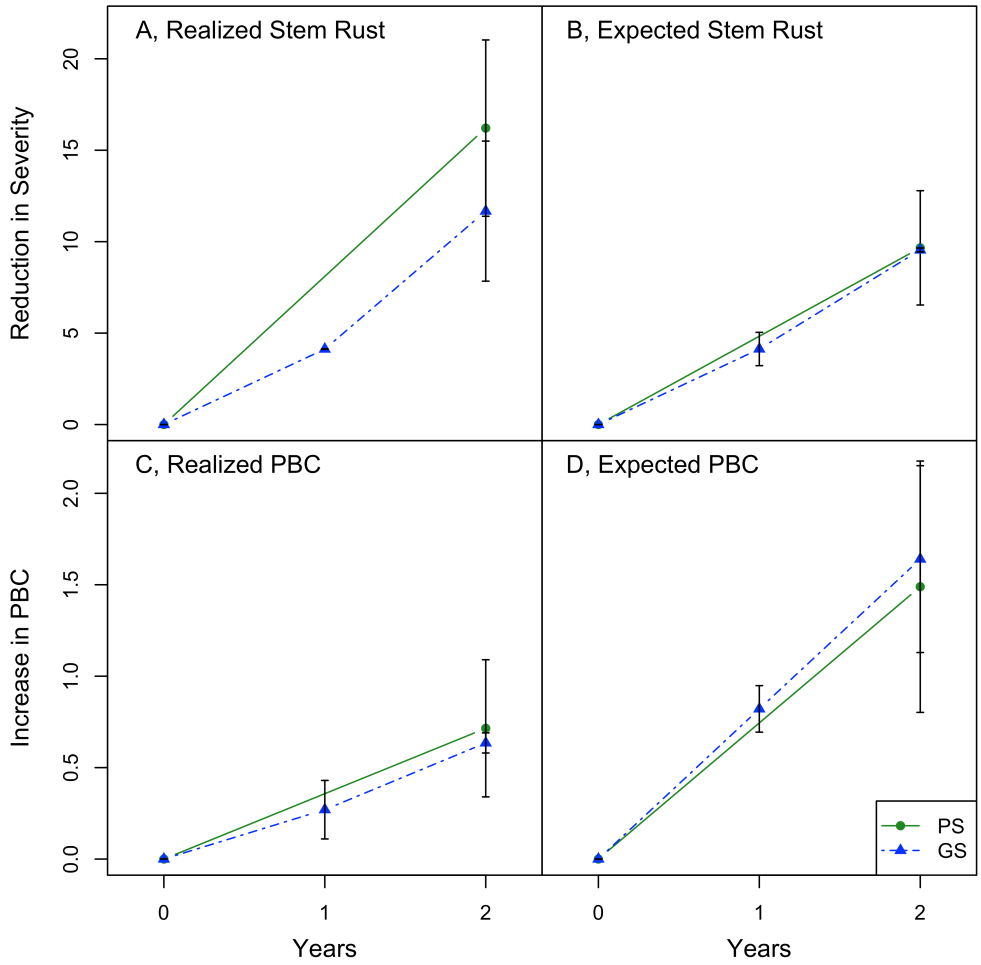


Figure 6.4: Realized and expected gain in stem rust resistance and PBC due to PS and GS for stem rust resistance. A, realized response for stem rust; B, expected response for stem rust; C, realized correlated response for PBC; D, expected correlated response for PBC. GS, blue triangles and solid lines; PS, green circles and solid lines. The x-axis indicates the year. One GS cycle requires one year and one PS cycle requires two years.

Realized and expected gain were not significantly different for PS cycle one, p=0.161 GS cycle one, p= 0.994 and GS cycle two, p= 0.67. Significant selection response was observed for both GS, p=0.018 and PS, p=0.00639 (Table 6.4).

Table 6.4: Rate of gain, significance of selection response, and percent total gain from GS and PS for stem rust resistance and PBC, a correlated trait.

	Selection treatment	Rate of gain	Standard error	T-value	P-value	Percent total gain
Stem rust resistance	GS	5.49	1.42	3.87	0.018	30.5±10.5
	PS	16.2	3.17	5.11	0.00639	41.9±11.8
PBC†	GS	0.307	0.0916	3.35	0.0285	252±35.9
	PS	0.714	0.205	3.48	0.0253	276±134

† PBC, pseudo-black chaff

Based on regression coefficients, rate of gain per cycle was 5.49±1.42 for GS and 16.2±3.17 for PS, corresponding to 14.3±3.69% gain per cycle for GS. Percent total gain was 30.5±10.5 and 41.9±11.8% for GS and PS respectively.

Because one cycle of PS required the same amount of time as two cycles of GS, we compared PS cycle one with GS cycle two to compare realized ΔG from both methods on a per unit time basis. For replicate one, realized ΔG from one cycle of PS was higher than realized ΔG from two cycles of GS, 21 *vs.* 7.84. In contrast, for replicate two, realized ΔG from one cycle of PS was lower than realized ΔG from two cycles of GS, 11.4 *vs.*15.5 (Table 6.2, Figure 6.5). Overall, realized ΔG from one cycle of PS and two cycles of GS was not significantly different, p= 0.692. Realized ΔG for PS and GS on a per-cycle basis was also not significantly different p=0.242.

Mean level of inbreeding and genetic variance

The inbreeding coefficients of C0, C1PS, C1GS, and C2GS, were

0.503±7.58e-4, 0.549±0.0215, 0.579±0.00767, and 0.714±0.0195 (Table 2). The change in inbreeding relative to C0 for C1PS, C1GS, and C2GS was 0.0462±0.0222, 0.076±0.00691, and 0.211±0.0187 (Figure 6.6A). The expected ΔF per cycle was 0.1. Actual ΔF per cycle was 0.0462±0.02 for PS and 0.0998±0.00895 for GS. Differences in ΔF between GS and PS were not significantly different per cycle, p=0.494, and per unit time, p=0.155.

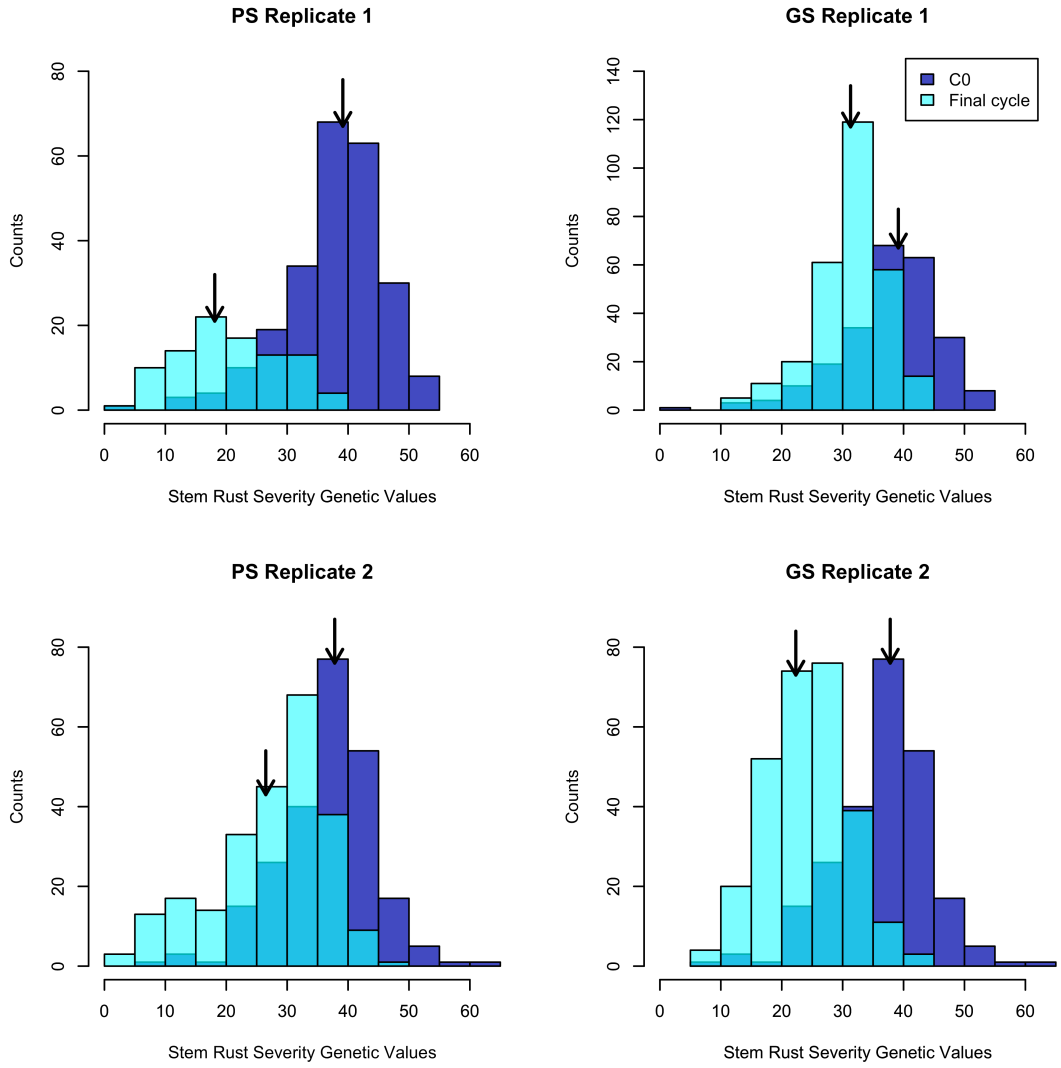


Table 6.5: Histograms of the genetic values for stem rust severity comparing C0 with the final populations from one cycle of PS or two cycles of GS. Adjusted population means are marked with arrows.

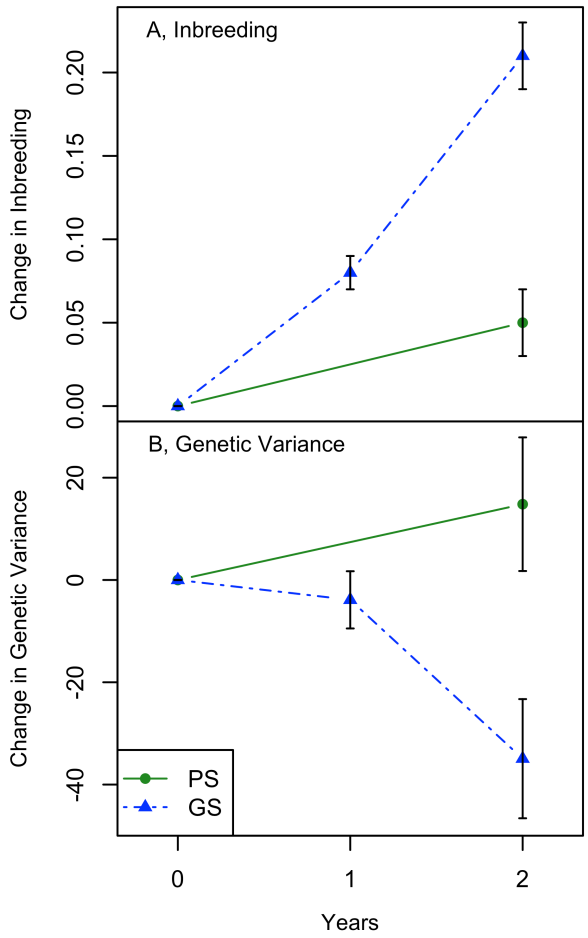


Figure 6.5: Change in inbreeding and genetic variance per year of GS and PS. A, Inbreeding; B, genetic variance; GS, blue triangles and solid lines; PS, green circles and solid lines. The x-axis indicates the year. One GS cycle requires one year and one PS cycle requires two years.

The genetic variance of C0 C1PS, C1GS, and C2GS, was 77.2±2.75, 92±10.3, 73.3±2.85, and 42.2±8.9 (Table 6.2). Change in genetic variance was 14.8±13.1, -3.88±5.6, -34.9±11.7 for C1PS, C1GS, and C2GS, respectively (Figure 6B). Per cycle change in genetic variance was not significantly different for GS and PS, p=0.242. However, on a per unit time basis, GS led to a significant reduction in genetic variance compared to PS, p=0.018.

Correlated response for pseudo-black chaff

Mean PBC of C0, C1PS, C1GS, and C2GS was, 0.265±0.0118, 0.966±0.383, 0.523±0.169, 0.885±0.0493. Expected and realized ΔG for PBC was 0.944±0.142 and 0.714±0.371 for PS cycle one, 0.821±0.127 and 0.27±0.158 for GS cycle one, and 1.64±0.722, and 0.633±0.0611, GS cycle two, respectively (Figure 4C,D). Realized and expected gain were significantly different for GS cycle one, p=0.0351, and not significantly different for PS cycle one, p=0.497 and GS cycle two, p= 0.267. Correlated response in PBC was significant for both GS and PS (Table 4), and based on regression coefficients, rates of correlated response in PBC from GS and PS for stem rust QR were 0.307±0.0916 and 0.714±0.205 per cycle respectively (Table 4), corresponding to a 121±36.2% per cycle gain from GS, and a 252±35.9, and 276±134% total gain from GS and PS respectively. ΔG for PBC from GS and PS were not significantly different per cycle, p=0.286, and per unit time, p=0.882.

Discussion

Effectiveness of selection

Significant gain from GS and PS was observed, and for both methods percent gain per cycle was quite high, 15.31±3.69 for GS and 41.93±11.8% for PS, suggesting that recurrent selection is a highly effective breeding strategy for stem rust QR. Other studies of recurrent selection for quantitative rust resistance in cereals have also reported high percent gain per cycle. For example, Díaz-Lago et al. (2002) achieved 11% gain per cycle from phenotypic recurrent selection for

partial resistance to crown rust (*Puccinia coronata* f. sp. *avenae)* in oats. In maize, recurrent selection for adult plant resistance to common rust (*Puccinia sorghi*) led to 28 and 6% gain per cycle in sweet (Abedon and Tracy, 1998) and tropical (Ceballos et al., 1991) germplasm respectively.

Expected and realized gain in stem rust quantitative resistance

Estimates of realized ΔG were in agreement with expected ΔG based on theory, however, there was a larger discrepancy between realized and expected ΔG for PS compared to GS. This could be due to drift, non-additive genetic variance that may have been exploited if the founders of C1PS contained favorable combinations at interacting loci, or due to rare recombination events between favorable alleles in negative linkage disequilibrium during the generation of the C1PS population.

Based on theory, we expect average ΔG from two cycles of GS to be higher than ΔG from one cycle of PS, leading to greater ΔG per unit time from GS compared to PS. However, in this experiment, due to the large standard errors of the mean expected and observed ΔG, we expected and observed that gain per unit time from GS and PS was not significantly different. Expected variation of selection response depends on the rate of inbreeding, $1/2N_e$ (Hill, 1977). The variation of selection response could have been reduced by increasing the number of replicates of the selection program, increasing N_e, using optimum contribution selection (Meuwissen, 1997; Grundy et al., 1998) to control the rate of inbreeding, or by optimizing selection to maximize response while

constraining its variance (Meuwissen and Woolliams, 1994).

Impact of selection on inbreeding and genetic variance

Although GS and PS both lead to significant response to selection, a slightly higher increase in mean inbreeding based on pedigree relationships was observed with GS compared with PS per unit time ($p=0.155$) largely due to the reduction in the breeding cycle duration. On a per cycle basis, inbreeding from GS and PS was similar and agreed with expected values. Simulation studies ignoring the impact of allele frequency changes have shown that inbreeding from GS is expected to be less than that of PS on a per cycle basis (de Roos et al., 2011; Dekkers, 2007) and greater than PS on a per unit time basis (de Roos et al., 2011). In a long term GS simulation study taking allele frequency changes into account, GS was shown to lead to greater genomic inbreeding rates both per cycle and per unit time (Jannink, 2010), and GS lead to a greater discrepancy between genomic inbreeding rates and pedigree based inbreeding rates. Long or medium term selection experiments will be needed to clarify how the implementation of GS will impact inbreeding rates in breeding programs.

A significantly greater reduction in genetic variance per unit time was observed for GS compared to PS. In selection programs, changes in genetic variation occur due to selection (Bulmer, 1971, 1980), finite population size, mutation, linkage between loci (Keightley and Hill, 1987) and the fixation of favorable alleles. With PS, selection can take advantage the additional variance due to new mutations and the increased genetic variance from rare

recombination events between favorable loci. In contrast, with GS based on markers only, selection cannot act upon sources of variance that arise from mutation and recombination as the selection candidates are generated. The larger per cycle reduction in genetic variance from GS compared to PS that we observed could have occurred because there were fewer sources of genetic variation that GS could exploit. The larger per unit time reduction in genetic variance from GS was due in large part to the reduction in the breeding cycle duration, but also may have been because GS can cause favorable alleles to be fixed more rapidly compared to PS (Jannink, 2010). Controlling inbreeding during GS by using optimum contribution selection based on genomic relationships (Sonesson et al., 2012) or by weighting low frequency favorable alleles in the genomic selection model (Jannink, 2010) could help reduce the loss in genetic variance from GS.

Correlated response in pseudo-black chaff

GS and PS for stem rust QR lead to a significant correlated response for PBC. This is expected because at least two loci affecting stem rust QR and PBC are known to be linked (Hare and McIntosh, 1979; Bariana et al., 2001; Yu et al., 2011; Singh et al., 2013). However, this is the first indication that PBC, scored on a zero to five scale, is associated with stem rust resistance. Percent gain in PBC was substantially higher than percent gain in stem rust QR, suggesting that the correlation between the two traits is due to relatively few loci. Expected correlated responses in PBC were always higher than observed responses, and in

the case of GS cycle one, expected and observed responses in PBC were significantly different. The discrepancy between observed expected responses may be due to, drift, inaccurate estimation of additive genetic variance, or genotype-by-environment interaction.

In order to produce germplasm with improved stem rust QR without increasing PBC to unacceptable levels, selection should be based on an index that includes both traits. Although with index selection, reduced gain from selection for each individual trait by $1/\sqrt{number\ of\ traits}$ is expected, it has been shown to be more effective than tandem or independent culling level selection (Hazel and Lush, 1942). If PBC is conferred by few loci, marker assisted selection could be an effective strategy, especially because expression of PBC does not occur in all environments.

Conclusion

This is the first comparison of realized gain from GS based on markers only with that of PS in crop plants. On a per unit time basis, responses in stem rust resistance and correlated responses in PBC was similar for both GS and PS. However, GS resulted in more inbreeding and significantly less genetic variance largely because of the reduction in the breeding cycle duration. Although recombination and mutation can act in each cycle to replenish the genetic variance, with GS, these forces were not sufficient to counter the reduction in genetic variance lost per cycle due to selection (Bulmer, 1971, 1980), finite population size, and linkage (Keightley and Hill, 1987). To achieve more

sustainable gains from GS in long or medium term breeding programs, optimum contribution selection (Meuwissen, 1997; Grundy et al., 1998) and weighting low frequency favorable alleles (Jannink, 2010) should be tested to help reduce the reduction in genetic variance due to inbreeding and the fixation of loci.

Overall gain from GS in this experiment may have been higher if the first cycle of selection had been based on phenotype as recommended by Bernardo and Yu (2007). This enables a relatively large improvement to be made initially due to high selection accuracy and high additive genetic variance. The initial population can then be used for model training for subsequent selection cycles. In this experiment, the high potential for gain from the first cycle of selection was not realized as a consequence of poor selection accuracy from the initial prediction model, which was trained using a relatively small number of historical individuals. After the model was updated with data from the initial population, GS accuracies doubled and more of the potential gain per cycle was realized.

For GS based on markers only to be worth the effort, expense, and potential reduction in genetic variance, it must outperform PS on a gain per unit time basis. The performance of GS relative to PS depends on the accuracy of selection, the selection intensity, and the breeding cycle time. Assuming that accuracy for GS based on markers only will always be less than that of PS, a substantial reduction in cycle time or a large increase in the selection intensity will be required in order to more than compensate for the accuracy reduction. In this experiment, the 50% reduction in cycle time from GS was not sufficient to

over-compensate for the loss in selection accuracy. The ideal marker only GS breeding scheme would allow the training population to be updated every selection cycle with accurate multi-location data, substantially reduce the breeding cycle time, and enable an increase in the effective population size and selection intensity.

Given the current status of prediction modeling, and genotyping technologies, GS based on markers only may not be advantageous or cost-effective for every crop-breeding program depending on its unique circumstances. GS based on genomic relationship and phenotypic information, as a way to increase the accuracy of selection rather than reduce the breeding cycle duration could be a more effective strategy in some cases. Individual breeding programs will need to empirically evaluate how best to use genome-wide marker technology in selection.

Acknowledgements

This research was funded by The Bill & Melinda Gates Foundation (Durable Rust Resistance in Wheat) and the United States Department of Agriculture[1]-Agricultural Research Service (USDA-ARS) (Appropriation No. 5430-21000-006-00D and Hatch 149-449). Partial support for J. Rutkoski was provided by a USDA National Needs Fellowship Grant #2008- 38420-04755 and an American Society of Plant Biology (ASPB) -Pioneer Hi-Bred Graduate Student Fellowship.

References

Abedon, B.G., and W.F. Tracy. 1998. Direct and indirect effects of full-sib recurrent selection for resistance to common rust (*Puccinia sorghi* Schw.) in three sweet corn populations. Crop Sci. 38: 56–61.

Asoro, F.G., M.A. Newell, W.D. Beavis, M.P. Scott, N.A. Tinker, and J.-L. Jannink. 2013. Genomic, marker-assisted, and pedigree-BLUP selection methods for β-glucan concentration in elite oat. Crop Sci. 53: 1894–1906.

Bariana, H.S., M.J. Hayden, N.U. Ahmed, J.A. Bell, P.J. Sharp, and R.A. McIntosh. 2001. Mapping of durable adult plant and seedling resistances to stripe rust and stem rust diseases in wheat. Aust. J. Agric. Res. 52: 1247–1255.

Bates, D., & Maechler, M. (2010). lme4: Linear mixed-effects models using S4 classes. Retrieved from http://cran.r-project.org/package=lme4.

Bernardo, R., and J. Yu. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 47: 1082–1090.

Box, G.E.., and D.R. Cox. 1964. An analysis of transformations. J. R. Stat. Soc. Ser. B 26: 211–252.

Bulmer, M.G. 1971. The effect of selection on genetic variability. Am. Nat. 105(943): 201–211.

Bulmer, M.G. 1980. The mathematical theory of quantitative genetics. Clarendon Press, Oxford.

Ceballos, H., J.A. Deutsch, and H. Gutiérrez. 1991. Recurrent selection for resistance to *Exserohilum turcicum* in eight subtropical maize populations. Crop Sci. 31: 964–971.

Crossa, J., G.D.L. Campos, P. Pérez, D. Gianola, J. Burgueño, J.L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger, and H.-J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713–724.

Dawson, J.C., J.B. Endelman, N. Heslot, J. Crossa, J. Poland, S. Dreisigacker, Y. Manès, M.E. Sorrells, and J.-L. Jannink. 2013. The use of unbalanced historical data for genomic selection in an international wheat breeding program. F. Crop. Res. 154: 12–22.

Dekkers, J.C.M. 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. J. Anim. Breed. Genet. 124: 331–341.

Díaz-Lago, J.E., D.D. Stuthman, and T.E. Abadie. 2002. Recurrent selection for partial resistance to crown rust in oat. Crop Sci. 42: 1475–1482.

Endelman, J. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Gen. 4: 250-255.

Federer, W.T. 2002. Construction and analysis of an augmented lattice square design. Biometrical J. 44: 251–257.

Flor, H.H. 1971. Current status of gene-for-gene concept. Annu. Rev. Phytopathol. 9: 275–296.

Gilmour, A.R., B.J. Gogel, B.R. Cullis, and R. Thompson. 2009. ASReml user guide release 3.0.VSN Intl. Ltd., Hemel Hempstead, UK.

Grundy, B., B. Villanueva, and J.A. Woolliams. 1998. Dynamic selection procedures for constrained inbreeding and their consequences for pedigree development. Genet. Res., Camb. 72: 159–168.

Hallauer, A.R., M.J. Carena, and J.B. Miranda Filho. 2010. Quantitative Genetics in Maize Breeding. Iowa State University Press, Ames, IA.

Hare, R.A., and R.A. McIntosh. 1979. Genetic and cytogenetic studies of durable adult-plant resistances in Hope and related cultivars to wheat rusts. Plant Pathol. 83: 350–367.

Hayes, B.J., P.M. Visscher, and M.E. Goddard. 2009. Increased accuracy of selection by using the realized relationship matrix. Genet. Res. (Camb). 91: 47–60.

Hazel, L.N., and J.L. Lush. 1942. The efficiency of three methods of selection. J. Hered. 32: 393–399.

Heffner, E.L., J.-L. Jannink, and M.E. Sorrells. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Gen. 4: 1–11.

Heffner, E.L., M.E. Sorrells, and J.-L. Jannink. 2009. Genomic selection for crop improvement. Crop Sci. 49: 1–12.

Hill W.G. 1977. Variation in response to selection. In: Pollak E., O. Kempthorne, T.B. Baily Jr. (eds) Proceedings of the International Conference on

Quantitative Genetics. Ames, IA, August 16–21, 1976, The Iowa State University Press: Ames, IA. pp. 343–365.

Jannink, J.-L. 2010. Dynamics of long-term genomic selection. Genet. Sel. Evol. 42:35.

Jin, Y., L.J. Szabo, Z.A. Pretorius, R.P. Singh, R. Ward, and T. Fetch. 2008. Detection of virulence to resistance gene Sr24 within race TTKS of *Puccinia graminis* f. sp *tritici*. Plant Dis. 92: 923–926.

Jin, Y., L.J. Szabo, M.N. Rouse, T. Fetch Jr, Z.A. Pretorius, R. Wanyera, and P. Njau. 2009. Detection of virulence to resistance gene Sr36 within the TTKS race lineage of *Puccinia graminis* f. sp. *tritici*. Plant Dis. 93: 367–370.

Keightley, P.D., and W.G. Hill. 1987. Directional selection and variation in finite populations. Genetics 117: 573–582.

Kempthorne, O., and R.N. Curnow. 1961. The partial diallel cross. Biometrics 17: 229–250.

Knott, D.R. 1982. Multigenic inheritance of stem rust resistance in wheat. Crop Sci. 22: 393–399.

Lorenz, A.J., S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata, K.P. Smith, M.E. Sorrells, and J.-L. Jannink. 2011. Genomic selection in plant breeding : Knowledge and prospects. Adv. Agron. 110: 77–123.

Massman, J.M., H.-J.G. Jung, and R. Bernardo. 2013. Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. Crop Sci. 53: 58–66.

McDonald, B. A., and C. Linde. 2002. Pathogen population genetics, evolutionary potential, and durable resistance. Annu. Rev. Phytopathol. 40: 349–379.

Meuwissen, T.H.E. 1997. Maximizing the response of selection with a predefined rate of inbreeding. J. Anim. Sci. 75: 934–940.

Meuwissen, T.H.E., and J.A. Woolliams. 1994. Response versus risk in breeding schemes. In: 5th World Congress on Genetics Applied to Livestock Production. Guelph, ON. 7-12 August 1994. University of Guelph, Guelph, ON. Vol. 18. pp. 236–243.

Nazari, K., M. Mafi, A. Yahyaoui, R.P. Singh, and R.F. Park. 2009. Detection of wheat stem rust (*Puccinia graminis* f. sp *tritici*) race TTKSK (Ug99) in Iran. Plant Dis. 93:317.

Ornella, L., S. Singh, P. Perez, J. Burgueño, R. Singh, E. Tapia, S. Bhavani, S. Dreisigacker, H.-J. Braun, K. Mathews, and J. Crossa. 2012. Genomic prediction of genetic values for resistance to wheat rusts. Plant Gen. 5: 136–148.

Park, R.F. 2007. Stem rust of wheat in Australia. Aust. J. Agric. Res. 58 : 558–566.

Parlevliet, J.E. 2002. Durability of resistance against fungal, bacterial and viral pathogens; present situation. Euphytica 124: 147–156.

Pérez, P., G. de Los Campos, J. Crossa, and D. Gianola. 2010. Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. Plant Gen. 3: 106–116.

Poland, J. A, P.J. Brown, M.E. Sorrells, and J.-L. Jannink. 2012a. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7: e32253.

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sánchez-Villeda, M. Sorrells, and J.-L. Jannink. 2012b. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Gen. 5: 103–113.

R Development Core Team. 2010. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/.

de Roos, A.P.W., C. Schrooten, R.F. Veerkamp, and J.A.M. van Arendonk. 2011. Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls. J. Dairy Sci. 94: 1559–1567.

Rutkoski, J., J. Benson, Y. Jia, G. Brown-Guedira, J.-L. Jannink, and M. Sorrells. 2012. Evaluation of genomic prediction methods for fusarium head blight resistance in wheat. 5: 51–61.

Rutkoski, J.E., E.L. Heffner, and M.E. Sorrells. 2010. Genomic selection for durable stem rust resistance in wheat. Euphytica 179: 161–173.

Sargolzaei, M., and H. Iwaisaki. 2005. Comparison of four direct algorithms for computing inbreeding coefficients. Anim. Sci. J. 76: 401–406.

Singh, R.P., D.P. Hodson, J. Huerta-Espino, Y. Jin, P. Njau, R. Wanyera, S.A. Herrera-Foessel, R.W. Ward, and L.S. Donald. 2008. Will stem rust destroy the world's wheat crop? Adv. Agron. 98: 271–309.

Singh, S., R.P. Singh, S. Bhavani, J. Huerta-Espino, and E.E. Lopez-Vera. 2013. QTL mapping of slow-rusting, adult plant resistance to race Ug99 of stem rust fungus in PBW343/Muu RIL population. Theor. Appl. Genet. 126: 1367–1375.

Sonesson, A.K., J.A. Woolliams, and T.H.E. Meuwissen. 2012. Genomic selection requires genomic control of inbreeding. Genet. Sel. Evol. 44:27.

Vazquez, A.I., D.M. Bates, G.J.M. Rosa, D. Gianola, and K.A. Weigel. 2010. Technical note: an R package for fitting generalized linear mixed models in animal breeding. J. Anim. Sci. 88: 497–504

Yu, L.-X., A. Lorenz, J. Rutkoski, R.P. Singh, S. Bhavani, J. Huerta-Espino, and M.E. Sorrells. 2011. Association mapping and gene-gene interaction for stem rust resistance in CIMMYT spring wheat germplasm. Theor. Appl. Genet. 123: 1257–1268.

CHAPTER 7

CONCLUSION

The overall goal of this work was to generate knowledge helpful for guiding the decisions of breeders and scientists working towards implementing genomic seleciton (GS) in wheat, especially for QR to Fusarium head blight (FHB) and stem rust. Before this work there were many unknowns about how to implement GS in wheat especially for quantitative resistance (QR). Genotyping-by-sequencing (GBS) was a new marker platform, and efforts to evaluate GBS for GS in wheat were just beginning. There was considerable anxiety about the high levels of missing data in GBS datasets, and many researchers were unsure how this would impact GS accuracy. There were also few options for missing data imputation in wheat because of the lack of an assembled genome sequence or sufficiently dense GBS map. Furthermore, the efficacy of GS for QR was considered questionable, and many believed that conventional marker assisted selection (MAS) strategies could be more effective. For model training, there was (and still is) considerable interest in using historical datasets, but studies to assess GS accuracy when using historical data for the prediction of new breeding materials had not been conducted in crops. Lastly, realized gain from GS had not been tested in wheat, and some wheat breeders were skeptical about how well GS could actually work in practice.

Before indicating how this work has contributed to the greater body of

knowledge, its important to emphasize that the conclusions presented here were drawn using data from few, or only one, dataset that may not be representative of other wheat data. Readers should confirm these findings using their own datasets prior to implementing GS in their programs.

This work has helped to fill several gaps in knowledge about how to implement GS. First, missing data is now known to have a minor impact on accuracy if marker densities are sufficiently high. For conservative GS practitioners who rely on un-ordered marker sets, random forest regression imputation (RFI) can be recommended prior to GS to ensure accuracy is not lost due to missing data. Second, GS, rather than conventional MAS can be recommended for QR, with the exception of the FHB resistance trait deoxynivalenol content which was best predicted using QTL linked markers alone. When performing GS for QR traits, loci targeted genotyping is recommended in addition to genome-wide genotyping because doing so could enable better modeling of major QTL. Third, model training with historical data has been shown to be risky. Results indicated that very large population sizes and high heritibilities will be required for historical data to achieve sufficient accuracies even when model training and validation occur within the same population. Lastly, GS for wheat stem rust QR has been shown to be as effective as phenotypic selection on a per-unit time basis, and realized gains were in agreement with expected gains, confirming that GS accuracy relative to phenotypic selection accuracy is a good indicator of the relative efficiency of GS.

This observation should also hold for other traits.

Several topics worthy of additional research have arisen as a result of this work. The development of faster unordered marker imputation methods that are at least as accurate as RFI on would be especially useful because the current implementation of RFI has a high computational burden. In addition, the value of loci-targeted genotyping to enable better modeling of major QTL should be confirmed for other traits where major QTL have been characterized. How to effectively use historical data for model training will require more studies specific to individual breeding programs and traits. In particular, the number of lines to phenotype and the number of evaluation environments for model updating will need to be determined to maximize genetic gain per unit time and cost. Finally, during the realized gain from GS study, GS was found to reduce genetic variance faster than phenotypic selection on a per unit time basis. The impact of GS on the genetic variance needs to be confirmed over more cycles of selection and for other traits, and the implications of this loss in genetic variance for medium term genetic gain will need to be further studied.